nature neuroscience

Article

https://doi.org/10.1038/s41593-024-01715-2

A nonoscillatory, millisecond-scale embedding of brain state provides insight into behavior

Received: 26 June 2023

Accepted: 19 June 2024

Published online: 15 July 2024

Check for updates

David F. Parks[1,6], Aidan M. Schneider[2,6], Yifan Xu[2], Samuel J. Brunwasser[2], Samuel Funderburk[2], Danilo Thurber[3], Tim Blanche[4], Eva L. Dyer[5], David Haussler[1] & Keith B. Hengen[2]

The most robust and reliable signatures of brain states are enriched in rhythms between 0.1 and 20 Hz. Here we address the possibility that the fundamental unit of brain state could be at the scale of milliseconds and micrometers. By analyzing high-resolution neural activity recorded in ten mouse brain regions over 24 h, we reveal that brain states are reliably identifiable (embedded) in fast, nonoscillatory activity. Sleep and wake states could be classified from 10^0 to 10^1 ms of neuronal activity sampled from 100 µm of brain tissue. In contrast to canonical rhythms, this embedding persists above 1,000 Hz. This high-frequency embedding is robust to substates, sharp-wave ripples and cortical on/off states. Individual regions intermittently switched states independently of the rest of the brain, and such brief state discontinuities coincided with brief behavioral discontinuities. Our results suggest that the fundamental unit of state in the brain is consistent with the spatial and temporal scale of neuronal computation.

For nearly a century, stereotyped electrical waves traveling across the surface of the brain have been used to define neural activity patterns correlated with sleep/wake state[1]. State-related waves are slow, travel across the isocortex and are detectable in anatomically distributed structures[2,3]. These waves require multiple seconds of observation for identification and can be measured through the scalp, which displays filtered activity averaged across many millimeters of isocortex[4]. In effect, state-related waves reflect a powerful mechanism of widely distributed electrical coordination, well poised to structure the global state of millions to billions of neurons over seconds to hours. As a result, the neural basis of sleep and wake states is generally understood to be a brain-wide phenomenon and orders of magnitude slower than the submillisecond precision of neural encoding of active behavior[5,6].

Contemporary studies of brain function and animal behavior are revealing complex state-related dynamics spanning multiple spatiotemporal scales, prompting calls for a reevaluation of traditional perspectives[7]. Specifically, there is significant heterogeneity in oscillatory activity within both sleep and wake[8–12]. The low-frequency waves that define sleep and wake travel across the isocortex but may be enriched or impoverished in different regions at any moment in time[13,14]. There is evidence that sleep and wake states may intrude on one another to some extent, even in the course of normal behavior[10,11,15–17]. The ability of the vertebrate brain to locally regulate states is exemplified by unihemispheric sleep in migratory birds[18], marine mammals[19] and potentially even humans[20]. Localized cortical states are also precedented in rodents[17,21]. This raises the question: what is the fundamental unit of brain state? A fundamental unit, in this context, is a minimal unit of neural activity which itself defines a brain state. A state's definition need not be a comprehensive characterization of all properties of the state but one that reliably distinguishes it from

[1]Department of Biomolecular Engineering, University of California, Santa Cruz, CA, USA. [2]Department of Biology, Washington University in Saint Louis, St. Louis, MO, USA. [3]Independent researcher, Exeter, NH, USA. [4]White Matter LLC, Seattle, WA, USA. [5]Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA. [6]These authors contributed equally: David F. Parks, Aidan M. Schneider. e-mail: khengen@wustl.edu

all other states in the set. One way to pursue answering this question is to ask, 'what is the minimal resolvable signature of the canonical states'? For example, if delta power (0.1–4 Hz), which is used to define slow-wave sleep, were the minimal unit of slow-wave sleep, examination of data in <1/4 s should carry no signature of the state.

The three broadest states, non-rapid eye movement (NREM) sleep, rapid eye movement (REM) sleep and waking, are composed of a hierarchy of substates and nested neurophysiological events, such as sleep spindles and sharp wave ripples. Additionally, the notion of discrete states is somewhat misleading, as transitions between states are characterized by intermediate behavior and neurophysiology[15,16].

Low-frequency (<100 Hz) dynamics are the foundation of previous descriptions of sleep/wake as well as emerging evidence of brief and localized state-related phenomena. These dynamics influence many levels of neuronal activity. For example, delta waves (0.1–4 Hz) originating in the prefrontal isocortex drive high-amplitude electroencephalography (EEG) waves that correlate with alternating periods of quiescence and bursting in isocortical neurons[22]. Bursting and silence is, unsurprisingly, supported by hyperpolarization and depolarization in neuronal membranes[3]. However, despite the fact that delta waves can be measured in a small patch of membrane, the fundamental unit of the wave is slow (250 ms to 10 s) and widely distributed, coordinating the activity of neurons over multiple centimeters[3]. As a result, local measurements of low-frequency waves are understood to reflect global, synchronizing forces[23,24]. This is the basis of a widely accepted model of brain states in which neuronal activity (fast and local) in distinct regions is systematically synchronized or desynchronized by oscillations (slow and distributed)[25,26].

Due to the physics of waves, it is impossible to consider a substrate of state shorter than the frequency of one cycle (and multiple cycles must be observed for practical purposes). As a result, frequency-based definitions of waves have a minimal resolution that is far slower and larger than the fundamental unit of neuronal activity: the action potential. A wave-based derivation that the fundamental unit of state is slow and global is, while perhaps true, tautological.

We sought to learn the minimally resolvable structure of sleep and wake directly from raw data, independent of assumptions. Convolutional neural networks (CNNs) are well suited to extract the rules of sleep and wake at different spatiotemporal scales. Crucially, CNNs function well with noisy label data—in practice, if a CNN is trained using human labels (based on waves) but learns a more reliable latent signature of state, it can overrule the training label and issue a high-confidence disagreement[27]. Using this bottom-up approach, we found that brain states are robustly resolvable in ≥40 ms of data from a single wire placed in any region in the brain. Removing oscillatory information below 750 Hz had no impact on accuracy. This suggests that the fundamental unit of brain state is at or below the spatiotemporal order of 10^1 ms and 10^2 µm. Fast and local embedding of state provides insights into brain function: individual regions intermittently switch states independently of the rest of the brain, correlating with brief behaviors in both sleep and wake.

Results

To empirically evaluate the minimally resolvable fingerprint of brain state in diverse regions throughout the mammalian brain, we analyzed a series of long-term, continuous multisite recordings in freely behaving mice. Briefly, each of nine included animals was implanted with multiple 64-channel microelectrode arrays. To facilitate stable, high signal-to-noise recordings of single-unit activity, arrays were composed of tetrodes. Arrays were attached to headstages by flex cables, allowing an arbitrary geometry of stereotactic targets. Recordings (~0.1 Hz to 7 kHz, sampled at 25 kHz) were conducted in the home cage continuously for between 4 and 16 weeks. Amplification, analog-to-digital conversion and multiplexing were done by the headstage (eCube HS-640, White Matter LLC), which was contained in a sparse three-dimensionally printed frame (Fig. 1a). The entire assembly weighed 4 g for a 512-channel (eight-module) implant. Neural signals were relayed to the data acquisition system (eCube Server, White Matter LLC) via a thin (1.0 mm) cable with a miniature inline commutator. High-resolution videos of behavior were collected in parallel with neural recordings (e3vision, White Matter LLC). Even in the case of an eight-site implantation, animal movement was similar to that of unimplanted animals (Supplementary Fig. 1).

Animals carried arrays in a minimum of three and a maximum of eight separate brain regions (Supplementary Fig. 2). Given the diversity of implantation geometries between animals, we ensured that each brain region included in this study was recorded in at least two animals. As a result, we examined a grand total of ten unique regions across these recordings (Fig. 1a): CA1 hippocampus (CA1), primary visual cortex (VISp), nucleus accumbens (ACB), primary somatosensory cortex (SSp), primary motor cortex (MOp), caudate/putamen (CP), superior colliculus (SC), anterior cingulate cortex (ACA), retrosplenial cortex (RSP) and lateral geniculate nucleus (LGN). Within each animal's dataset, we arbitrarily selected a 24 h period for analysis, ensuring only that there was high-quality electrophysiological data in every brain region (that is, evidence of single-unit activity; Supplementary Fig. 3 and Supplementary Table 1). Note that, while we used the presence of single-unit activity to indicate high-quality data, our analyses utilized the broadband signal (raw data) unless otherwise indicated (Fig. 1b).

Three human experts independently sleep scored the 24 h datasets, labeling waking, REM and NREM sleep using polysomnography (Extended Data Fig. 1). The three experts then met to address points of disagreement in each dataset, thus generating a consensus score that served as the basis of subsequent comparison. Substates within the three states, such as active/quiet wake and sleep spindles, were identified algorithmically and confirmed by human scorers.

CNNs extract arousal states from brief, local observations

Sleep scoring is traditionally conducted taking advantage of EEG, which reflects electrical waves on the dorsal isocortical surface, albeit with low spatiotemporal resolution. Traditional sleep scoring requires some measure of animal motor output, such as electromyography (EMG). While neural activity throughout the brain is influenced by sleep and wake[15,28,29], the degree to which sleep and wake are discernable solely on the basis of these dynamics is less explored. To test this, and allow for region-by-region variability in state-related dynamics, we trained and tested a unique CNN on the broadband raw data recorded in each region (Fig. 1c). Summarily, provided with only locally recorded raw neural data, the CNN attempts to learn rules that consistently predict human-generated labels, even though those labels were derived from EEG and behavior (Fig. 1d).

Each CNN was shown 2.6 s of data at a time (2.6 s CNN) and was trained on 18 h of data and tested on a withheld, contiguous 6 h block of data that spanned a light–dark transition (Extended Data Fig. 2). In all animals, the withheld test set was temporally separated from the training set, ensuring some generalization of the model into the past and future. Despite the fact that CNNs only observed data from an individual region and had no information about animal movement, all three states (REM, NREM and wake) were robustly separable in 2.6 s increments of data from every region. The accuracy was comparable to that achieved by human experts (Fig. 1d–f). CNNs effectively detected brief states, such as microarousals, and in some cases identified examples missed by individual human scorers (Extended Data Fig. 1f).

CNNs can learn complex patterns from raw data[27], but understanding the rules learned by models is a widely recognized challenge[30]. One approach to this is ablation: by removing key components of a dataset, one can ascertain whether those components are integral to a model's success. Because human experts rely on low-frequency patterns (0.1–16 Hz (ref. 31)), we hypothesized that CNNs were using the same information. To test this, we applied a series of band-pass

https://doi.org/10.1038/s41593-024-01715-2

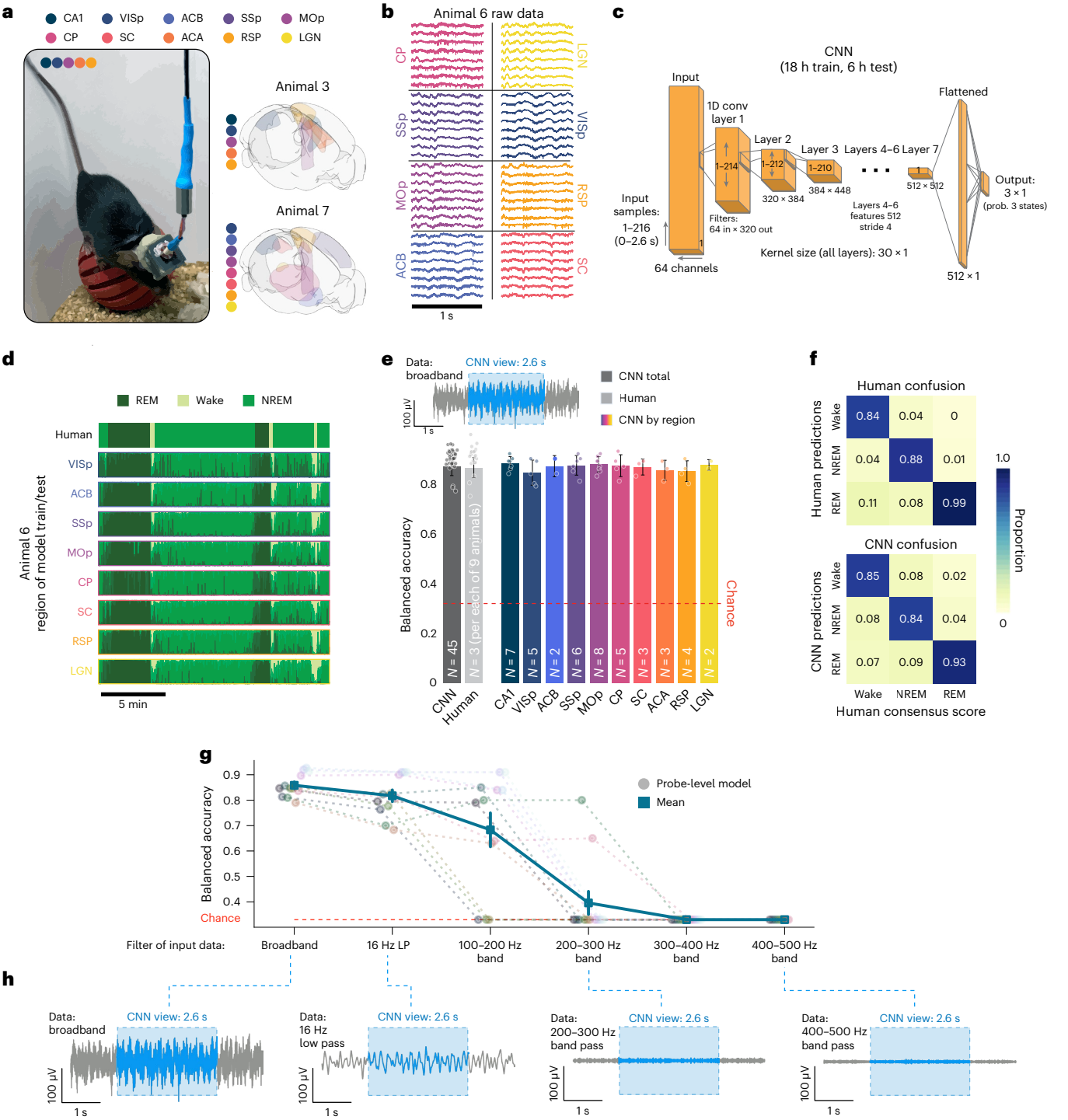


Fig. 1 | CNNs learn robust signatures of sleep and wake from raw neural data in all brain regions. a, Recording protocol. Left: image of a freely behaving mouse carrying a continuous multisite recording device. Right: Brainrender examples of implant/recording geometry from two of the nine animals in this study[71]. The colored regions indicate the recorded regions. **b**, One second of raw data from 8/64 channels in each of eight implanted brain regions in animal 6. **c**, The architecture of the convolutional neural network (CNN) used to decode brain state (wake, REM and NREM sleep) from raw data (Methods). Diagram shows convolutional (conv) and feedforward layers. **d**, Human scoring of state (top) versus eight CNNs (bottom) trained and tested independently on eight brain regions in the same animal. The *y* axis of each row represents the probability (prob.) of each of the three states over time. **e**, The CNN accuracy relative to a consensus score is slightly but significantly better than that achieved by individual human experts (left; gray bars, $P = 0.011$, one-way ANOVA). Colored bars: the CNNs trained and tested in each of ten brain regions show comparable accuracy; the text in the bars indicates *n* animals in each region ($P = 0.180$, one-way ANOVA). Top inset: 5 s of raw data (gray) and a 2.6 s sample (blue) that the CNN uses for classification. **f**, Confusion matrices comparing human scorers (top) and CNNs (bottom) against consensus scores. Human scorers utilize full polysomnography. CNNs achieve balanced results across three states by observing only raw neural data. **g**, To test the source of state information learned by the CNN, models were trained and tested on filtered raw data from a subset of probes (12 regions from two animals): broadband (unaltered raw data), low-pass filtered at 16 Hz, and a series of progressively higher band-pass filters. The balanced accuracy of models is shown as a function of filter. **h**, Visual summary of the filters applied (gray shows the same 5 s of data in each example; blue is the 2.6 s window visible to the CNN for scoring). The data are presented as mean ± s.e.m.

https://doi.org/10.1038/s41593-024-01715-2

filters to data from each region in two animals (total of 12 regions); after filtering, only a specific range of frequencies remained. In each region, a new series of 2.6 s CNNs was trained and tested on each of five band-passed datasets: (a) 0.1–16 Hz, (b) 100–200 Hz, (c) 200–300 Hz, (d) 300–400 Hz and (e) 400 –500 Hz (Fig. 1g,h). Canonical state information is richly embedded in the 0.1–16 Hz band, but gamma power (30–100 Hz) changes by state as well. If CNNs required canonical oscillatory information, we expected accuracy to decline across these filters. Consistent with this, the 0.1–16 Hz model performed within 3% of broadband models (Fig. 1g). CNNs were progressively impaired by the remaining band passes; all models reached chance levels by 300–400 Hz. Taken together, these data demonstrate that CNNs can robustly identify REM, NREM and wake in only 2.6 s of raw neural data from any recorded region and that, consistent with nearly a century of observation[1], models rely on low-frequency information to achieve this (Fig. 1h).

State structures activity at the kilohertz and 10^1 ms scale

The lowest band pass suggests that complete information about states is available at frequencies below 16 Hz, and the progressive band passes confirm that state information declines to 0 by 200 Hz. However, these results do not rule out the possibility that a mechanistically distinct source of state embedding might emerge in ultrafast frequencies. We sought to rule this out by training 1 s CNNs on datasets subjected to a series of increasing high-pass filters. In other words, we eliminated low-frequency information progressively, allowing all faster information through at each step. Consistent with our prior results, high-passing at 0.3 Hz did not affect model accuracy. Similarly, high-passing above the delta band (4 Hz) had no impact on accuracy. Unexpectedly, this trend continued across almost four orders of magnitude (10^{-1} to 10^3): models maintained full accuracy at 750 Hz and dropped to 0 only at >3,000 Hz (Fig. 2a). Models failed in a stepwise fashion: accuracy generally remained above 70% before dropping to chance (Extended Data Fig. 3a). The 1 kHz range is understood to contain only fast neuronal events such as action potentials[32]; extracellular action potentials disappeared between 1,000 and 5,000 Hz (Fig. 2a).

Across regions, 1 s CNNs trained on 750 Hz high-passed data matched the performance of baseline (2.6 s broadband CNNs) and low-frequency models (1 s low-pass CNNs) (Fig. 2b). Alongside the failure of band-pass models between 200 and 500 Hz, these data strongly suggest that brain state is embedded in spike-band frequencies. Despite being separated from traditional metrics by multiple orders of magnitude, models restricted to kilohertz-range data suffer no loss in accuracy.

A trivial explanation is that spike-band information simply reflects low-frequency rhythms: action potential bursting and silence are shaped by low-frequency field potentials[17]. Put simply, high-pass CNNs may be learning to reconstruct slow information in the timing of fast events. To investigate this, we took advantage of the timescale of oscillatory information: a 1 Hz wave cannot fit in windows smaller than 1 s. We trained another series of CNNs, this time systematically manipulating input length. We progressively reduced input size from 2.6 s down to a single sample point, or 1/25,000 s (Fig. 2c). Each model was provided with data from all channels within an individual region, limited to a given input size. While median model performance declined monotonically as a function of input size, CNNs maintained accuracy significantly above chance down to 5 ms (0.43 ± 0.02, chance 0.33). Given only 40 ms of input, median CNN accuracy was nearly 70% (0.68 ± 0.02, chance 0.33), which suggests that the average 40 ms model was capable of learning something about each of the three states. Note that chance is precisely 0.33, CNNs are provided with balanced datasets, and the mode of failure is characterized by universally guessing a single class (for example, NREM). We excluded models that failed in this respect from statistical comparison. Interestingly, we observed relatively balanced performance across the three states all the way down to 5 ms (Fig. 2d); in other words, CNNs did not struggle to identify one of the states relative to the other two. Taken with the high-pass results, these data suggest that all three states impose discriminable structure on neural activity at the millisecond timescale.

We next reasoned that CNNs may be learning to reconstruct slow waves in short intervals by virtue of spatial information gleaned by sampling 64 channels. Succinctly, because a wave travels in space, an instantaneous measurement of voltage across multiple electrodes could provide a measure of wavelength. To address this, we trained another series of CNNs exposed to progressively smaller and smaller input sizes. This time, however, we only passed in data from a single channel. While overall accuracy was slightly reduced and more variable in this heavily constrained situation, we observed qualitatively similar results in all single-channel models from all regions of all animals (Extended Data Fig. 3b). Many single-channel models performed significantly above chance down to 1 ms.

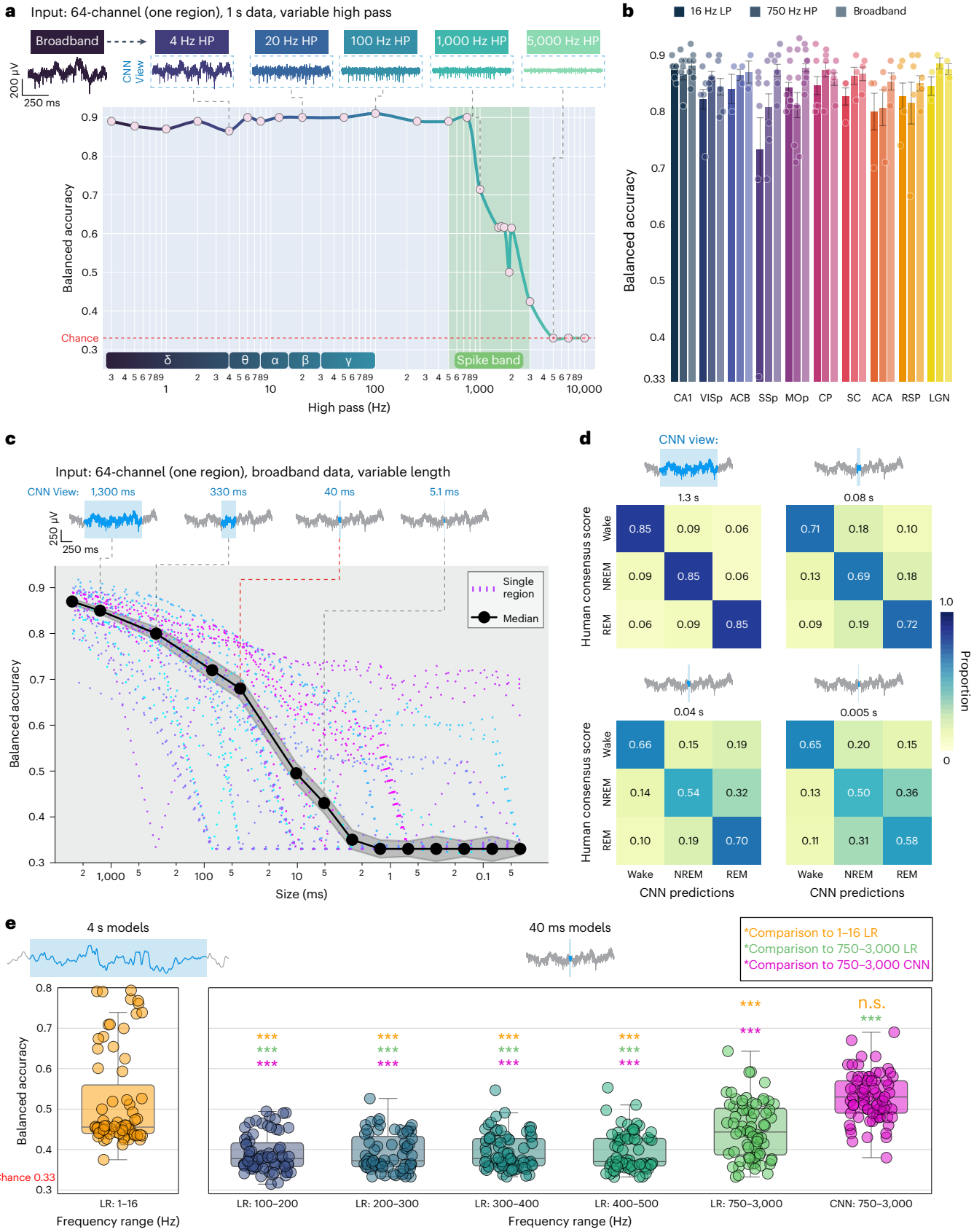
Historically, neurophysiological definitions of sleep and wake center around state-related changes in spectral density power. Thus, a plausible source of fast state-embedding could be systematic changes in high-frequency bandpower. We addressed this directly by extracting spectral density data using a fast Fourier transform (FFT) applied to a series of band-pass-filtered signals. These spectral density data, which represent the power distribution across different frequency bands, were then used to train a logistic regression (LR) model to classify states.

To establish a baseline, we first focused on the 1–16 Hz frequency band. We passed single-channel data in 4 s increments, each subjected to FFT, into our LR model. This was intended to validate the model's capacity to discern brain state in a well-studied frequency range. In this context, LR achieved a mean balanced accuracy of 51.68 ± 1.5% (Fig. 2e).

Fig. 2 | Brain state can be recovered in the kilohertz band and only 10^0–10^1 ms of data. a, Top: 1 s of raw neural data from all channels in one animal subjected to increasing high-pass filters up to 10,000 Hz. Note that spikes are eliminated between 1,000 and 5,000 Hz. Bottom: CNN performance as a function of increasing high-pass filters in an example animal. High-pass filtering only decreased brain state information above 1,000 Hz ($y = -9.892 \times 10^{-5}x + 0.84$, $P < 2 \times 10^{-16}$, $r^2 = 0.43$). Bottom inset: depiction of classical oscillatory bands used to define states. **b**, Bar plot summarizing model accuracy by region in three conditions: 16 Hz low-passed (LP) data, 750 Hz high-passed (HP) data and broadband (unfiltered) data from all animals ($n = 9$ mice incorporating 45 probes in total). Broadband accuracy was slightly but significantly higher than low-pass accuracy and high-pass accuracy ($P = 0.020$ and 0.014, respectively; linear mixed effects: model accuracy ~ filter × region + (1|animal), where animal is a random effect) (EMMeans with Tukey post hoc correction). The data are presented as mean ± s.e.m. **c**, Top: 1 s raw data (gray) overlaid with progressively shorter CNN input sizes (blue). CNNs observe single input segments at a time for classification. Bottom: balanced accuracy of CNNs trained and tested on progressively reduced input size. Each recorded region ($n = 45$ implants, 9 animals, 10 regions) is plotted individually (dashed lines), with the overall median performance illustrated (solid black line). The data are presented as mean ± s.e.m. **d**, Confusion matrices of mean performance of all models above chance at four input sizes: 1.3 s, 80 ms, 40 ms and 5 ms. The values and colors represent class-balanced accuracy. **e**, The state-related content of six nonoverlapping frequency bands is assessed by logistic regression (LR) of the fast Fourier transform (FFT). As a positive control, 4 s of single-channel data were band-passed at 1–16 Hz (left, gold). Middle: LRs classified state in 40 ms segments of data in five band passes: 100–200 Hz, 200–300 Hz, 300–400 Hz, 400–500 Hz and 750–3,000 Hz. Right: for comparison, the 750–3,000 Hz/40 ms single-channel CNN is shown in magenta. For scale, 40 ms of data and 4 s of data are shown in blue above the relevant plots. The box plots indicate intermediate quartiles, and whiskers extend 1.5 interquartile ranges. The colored swarms indicate models trained on individual channels. $n = 71$ individual channels, where each implant is represented at least once. EMMeans with Tukey post hoc. n.s., $P > 0.05$. *$P \leq 0.05$, **$P < 0.01$, ***$P < 0.001$.

https://doi.org/10.1038/s41593-024-01715-2

We then performed the same analysis on 40 ms segments of data subjected to five increasing band passes: (a) 100–200 Hz, (b) 200–300 Hz, (c) 300–400 Hz, (d) 400–500 Hz and (e) 750–3,000 Hz (Fig. 2e and Extended Data Fig. 4). As expected, accuracy dropped to near chance between 100 and 500 Hz, significantly below the 1–16 Hz band (a: 38.98 ± 0.58%; b: 39.09 ± 0.57%; c: 39.43 ± 0.56%; d: 39.46 ± 0.60%). To our surprise, however, classifier accuracy increased significantly in the 750–3,000 Hz band (e: 44.77 ± 0.88%). Examination



https://doi.org/10.1038/s41593-024-01715-2

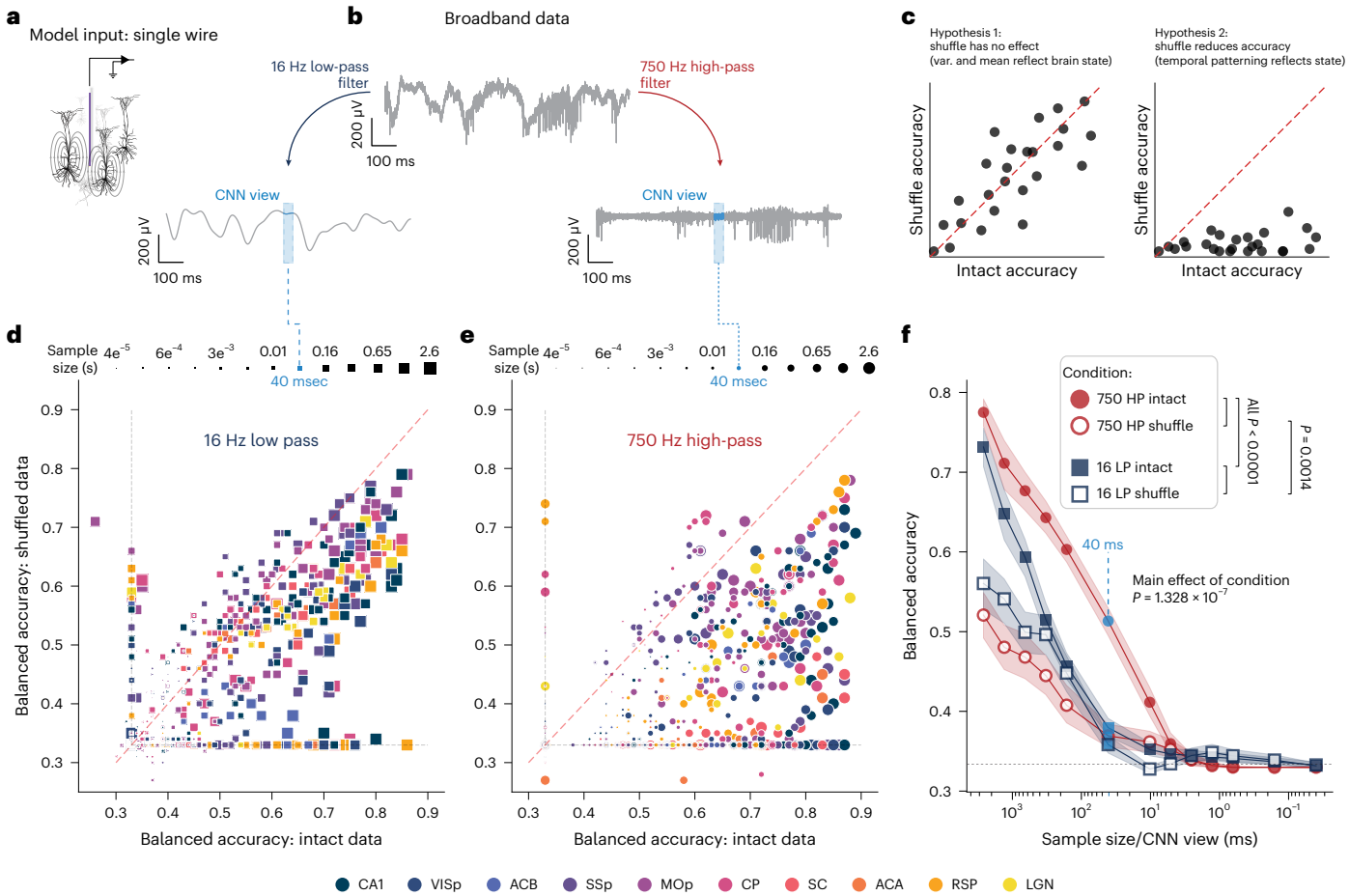


Fig. 3 | Brain states pattern high-frequency neuronal dynamics on the order of 10^0–10^1 ms. a, Experimental overview: model input. Depiction of a single wire (m) placed in a local region. The curved lines represent the spatial effects of single-neuron current dipoles that influence measured voltage at high frequencies. **b**, Experimental overview: paired low-pass and high-pass models. Parallel models are trained and tested on data from the same single channel (A), one model observing 16 Hz low-passed data, the other observing 750 Hz high-passed data. The blue boxes depict a 40 ms observation interval in each case. **c**, In each condition (low- and high-pass), another pair of models are made: intact and shuffled data (each sample is shuffled before training/testing). Left: if shuffling does not reduce the ability of a CNN to decode state, state information must be recoverable from sample mean and variance. Right: if temporal pattern is determined by state, shuffling will reduce accuracy. **d**, 16 Hz low-passed single channel data. Each square shows a pair of intact/shuffled models. The square is colored by the region on which they are trained, and the size of the square indicates input size. **e**, The same as **d**, but for 750 Hz high-passed single channel data. In **d** and **e**, two single channels were selected in each recorded region (both with and without high-amplitude spiking) for examination in four conditions: high-/low-pass filtering and shuffle/intact comparison. In each condition, models were trained/tested at 13 input sizes (shown along the top) from 2.6 s (65,536 data points) to 0.04 ms (1 data point), yielding a total of 2,028 single-channel CNNs. **f**, Summary of models in **d** and **e**. Red indicates 750 Hz high-pass (HP), and blue indicates 16 Hz low-pass (LP). The data are presented as mean ± s.e.m. The filled points are intact data, and the open points are shuffled samples. For ease of comparison with other results, 40 ms is indicated by a dashed light-blue line. Linear mixed effects balanced accuracy ~ input size × filter × shuffle + (1|animal) + (1|region). EMMeans with Tukey post hoc correction was used for statistics.

of the weights learned by LR models revealed an interesting phenomenon: 1–16 Hz models recapitulated established heuristics, for example, delta power contributes to NREM identification, and theta contributes to REM. Above 750 Hz, however, frequency-based patterns were unique to each model, suggesting that signatures of state in this range were local (Extended Data Fig. 4d,e).

States impose millisecond-scale patterning in neuronal activity

We next sought to understand how state is embedded in neuronal activity at high frequencies and in small intervals. Logically, there are three possible mechanisms by which this could arise. First, models could learn that average instantaneous voltage differs by state. This is the only possible explanation of performing above chance when testing on a single sample point. Second, variance might differ by state. In this case, two or more sample points could provide insight. Each of these explanations is decodable in fast/small samples but does not necessitate a fast underlying process. Third, neural activity could be patterned at the millisecond timescale. In this case, the sequence of voltage measurements carries state information and requires a fast organizing process.

To disambiguate these options, we trained a series of single-channel CNNs (Fig. 3a) on the same datasets in two conditions. One was low-passed at 16 Hz, and one was high-passed at 750 Hz (Fig. 3b). In effect, after high- and low-pass filtering, the same sample was passed into corresponding CNNs for labeling. In these parallel high- and low-pass datasets, we trained a series of CNNs on systematically decreased sample intervals. Finally, to test the hypothesis that high-frequency state embedding might reflect patterned information, we trained another parallel set of models but shuffled each sample before input (Fig. 3c). To summarize, progressively smaller chunks of data from a single wire were sleep scored by CNNs in four scenarios: intact low-pass, intact high-pass, shuffled-low pass and shuffled-high pass.

https://doi.org/10.1038/s41593-024-01715-2

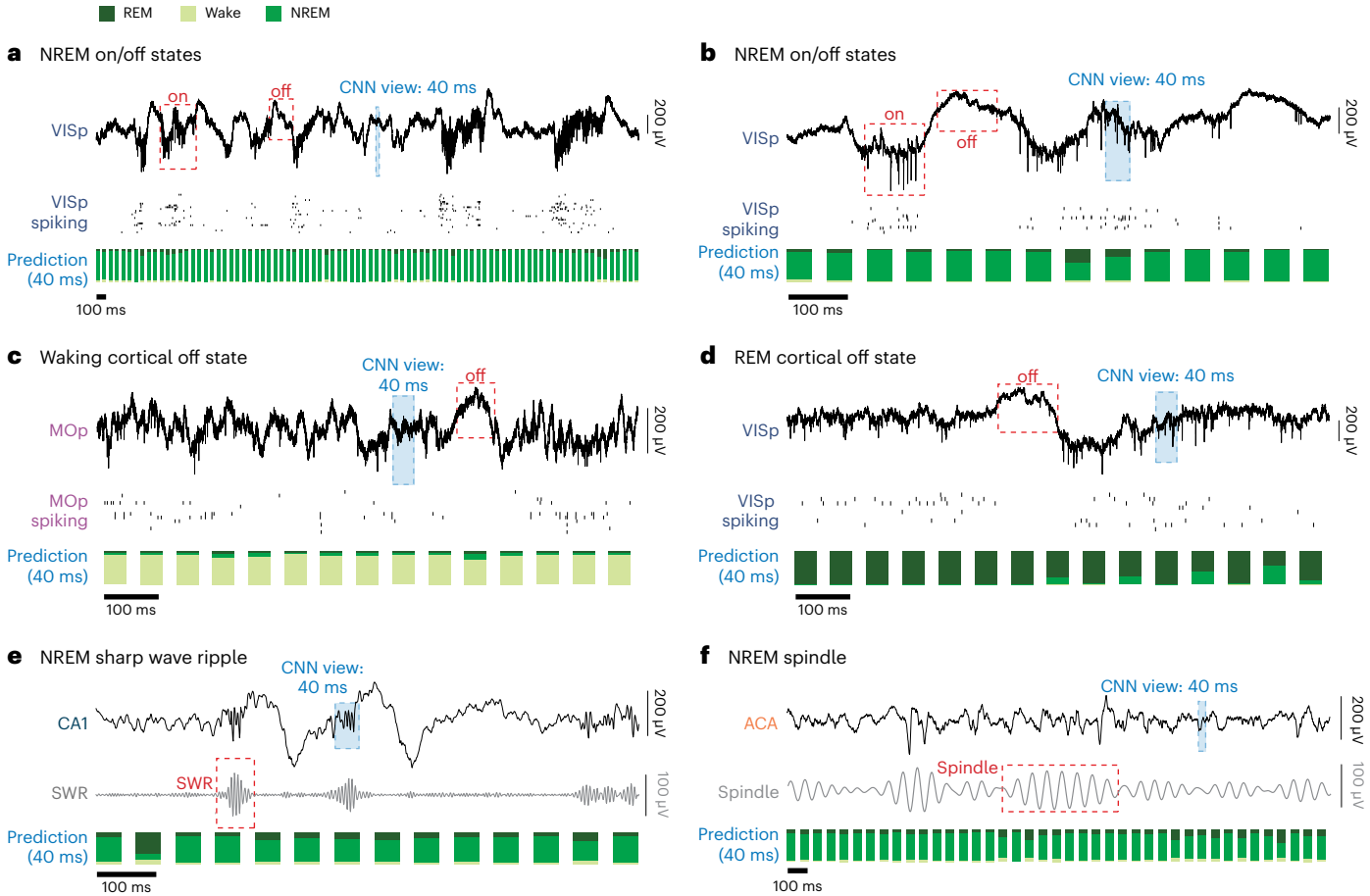


Fig. 4 | Fast embedding of states is robust to diverse low frequency activity and neurophysiological events. a, Top: broadband trace of several seconds of exemplary high delta (0.1–4 Hz) activity during NREM sleep. The data are recorded in VISp. The red boxes indicate cortical on and off states. The blue box shows the width of an individual input sample used by the 40 ms CNN to predict state. Middle: raster of subset of VISp single units spiking. Bottom: stacked bar plot of 40 ms CNN prediction probabilities (the three colors in each bar show the probability that the corresponding sample came from each of the three states). To reduce computational burden, the CNN evaluates a 40 ms sample every 1/15 s (hence the slight gaps between samples). **b**, Zoomed 1 s view of on/off states in VISp. **c**, Example of a waking off state in MOp. **d**, Example of a REM off state in VISp. **e**, Example of a NREM sharp wave ripple (SWR) in CA1 hippocampus. The middle trace shows the same data as the top trace (LFP, <200 Hz) but filtered to highlight SWR. **f**, Example of a NREM spindle in ACA. The middle trace shows the same data as the top trace (LFP, <200 Hz) but filtered to highlight spindles.

In the time intervals examined (2.6 s to 1/25,000 s), intact low-pass models performed only slightly better than their shuffled counterparts. This indicates that, at low frequencies, temporal organization of data is not necessary to estimate state. A comparison of paired intact/shuffled lowpass models revealed a tendency to cluster near the unity line (Fig. 3d). The performance of shuffled and intact low-pass models was only separable down to 655 ms ($P < 0.0001$; linear mixed model: balanced accuracy ~ sample size × high low pass × shuffle + (1|animal) + (1|region), where animal and region are random effects), after which they converged and then drop to near chance (<40%) by 40 ms (Fig. 3f). This confirms the efficacy of the low-pass filter: slow information cannot fit into short windows. These data demonstrate that low frequencies carry some information about brain state at relatively short intervals but that this information is distributional; that shuffling does not impair performance reveals that models are learning from mean and/or variance but not pattern.

In contrast, intact 750 Hz high-pass models significantly outperformed their shuffled analogs at all sizes above 2.5 ms. A comparison of paired intact/shuffled models reveals a rightward shift toward intact models (Fig. 3e), demonstrating that high-frequency/fast state information is embedded in temporal patterning (Fig. 3c). Not only did intact 750 Hz high-pass models outperform their shuffled variants, they also significantly outperformed all other models above 2.5 ms, including intact low-pass models at 2.6 s ($P = 0.0036$). At 40 ms, intact high-pass models was the only models to perform above 40% (near chance), yielding an accuracy of 51.3 ± 0.9%. This is consistent with the results of 64-channel CNNs tested on broadband data (Fig. 2c), which suggested that 40 ms is the minimal sample size that maintains high accuracy, at least within the limitations of our recordings. Taken together, these data reveal that patterns >750 Hz carry state-related information on the timescale of 5 ms and longer. Due to the effects of space filtering, it is likely that these patterns arise within <100 µm of each recording site[33] (Extended Data Fig. 4e).

Surprisingly, the effect of shuffling was observed similarly on channels both with and without high-amplitude spiking. That even nonspiking channels outperformed their shuffled counterparts reveals that the background recording 'noise' carries temporally structured information about brain state. A possible source of this information is low-amplitude spiking of nearby neurons, that is, multiunit hash[34,35].

Substates and fast neurophysiological events

Sleep and wake comprise many substates. In addition, stereotyped, intermittent neurophysiological events unfold in a state-dependent fashion. Sharp wave ripples are enriched in NREM sleep and quiet waking[36,37], while cortical on and off states are primarily associated with NREM sleep[17]. These examples raise two questions about short and

fast embedding of brain states. First, could it be that CNNs rely on fast substates and events (for example, sharp wave ripples) to achieve their accuracy? This is unlikely because fast events such as sharp wave ripples comprise a tiny fraction of total time in a state. Second, given that key neurophysiological events are defined by rapid and dramatic changes in activity, do such events lead to model errors?

To address these questions, we used established algorithms to detect sharp wave ripples[38,39], sleep spindles[40] and cortical on and off states[17]. In addition, we divided all of waking into two substates, active and quiet wake. We then evaluated the accuracy of 40 ms CNNs in and around these events.

Cortical on and off states during NREM did not confuse 40 ms models, despite the fact that individual on and off states lasted longer than 40 ms (Fig. 4a,b and Extended Data Fig. 5e). This suggests a reliable latent signature of NREM that supersedes an off state. Cortical on and off states also occur in waking and REM, albeit less frequently[17,41]. To a human observer, 40 ms of data within an off state from REM, NREM and wake are identical, yet 40 ms CNNs correctly identified the superstate of these on and off states (Fig. 4c,d and Extended Data Fig. 5e). Likewise, the 40 ms CNN was robust to sharp wave ripples (Fig. 4e and Extended Data Fig. 5e) and sleep spindles (Fig. 4f and Extended Data Fig. 5e). Finally, network activity differences between quiet and active wake did not drive confusion (Extended Data Fig. 5a,b). Given the absence of obvious neuronal activity during off states, it is likely that the background multiunit hash (that is, the 20 μV 'noise' band) carries meaningful information about state in the form of temporal patterning (Fig. 3).

Individual regions exhibit brief, independent states

We noted that CNNs of all sizes occasionally gave rise to high-confidence errors and even disagreed with labels in training datasets (Extended Data Fig. 6). Interestingly, these disagreements often spanned independent models trained in distinct regions in the same brain. Given the confidence and independent reproducibility of these events, we reasoned that they may be evidence of two known phenomena. First, microstates have been described extensively[42]. Microarousals are brief (3–15 s) global transitions from NREM sleep to wake[42,43] followed by a prompt return to NREM.

Microsleeps are equivalent but involve a brief intrusion of sleep into waking. Second, slow waves characteristic of sleep may appear in portions of the cortex of awake animals in conditions of sleep deprivation, inactivity or inattention[17,44]. This is referred to as local sleep.

We operationally defined microarousals as brief periods (<20 s) occurring during otherwise consolidated sleep in which independent 1 s CNNs in each region identified waking with high confidence (Fig. 5a, left). We applied an equivalent definition to microsleeps, identifying periods during waking when all models briefly identified a sleep state (Fig. 5a, middle). We eliminated these from our analyses. This excluded roughly 10% of NREM-to-wake disagreements and 5% of wake-to-NREM disagreements. After this, we were left with brief states that occurred in only a subset of regions (that is, not global), which we termed 'flickers' (Fig. 5a, right).

We initially hypothesized that flickering reflected local sleep[17]. To evaluate this, identified flickers in the output of 1 s CNNs operating on 16 Hz low-passed data. Approximately 60% of flickers indicated in the broadband model were captured by low-pass models (Fig. 5b). These were consistent with human readable, low-frequency local states, such as slow waves during wake[17] and REM[41], local changes in slow waves during NREM[45] and local theta power during patterned behavior[8,9]. We eliminated these from our analyses. We also directly evaluated whether flickers represent confusion driven by episodic oscillatory events. Neither on/off states, sleep spindles nor sharp wave ripples correlated positively with flickers. We reasoned that remaining events—those not visible to low-pass models—might represent transient, local shifts in the high-frequency latent patterning of state.

Note that >99% of excluded global events were also detected by low-pass models. This supports the efficacy of, for example, EEG, for the detection of microarousals and microsleeps.

To test the sensitivity of CNNs to momentary switches of state, we created a positive control. We randomly transposed variable length snippets of data between stretches of REM, NREM and wake (Fig. 5c). The 655 ms CNNs detected such synthetic flickers down to 10 ms and detected all examples ≥133 ms. This demonstrates the capacity of CNNs to detect brief intrusions of one state into another.

Because flickering appeared to be a plausible neurobiological event, we systematically quantified flicker rate and duration as a function of region. Flickers occurred at a rate of 10–50 h^{-1} in each region (Fig. 5d, $P = 1 \times 10^{-22}$, main effect of region, linear mixed model: flicker rate ~ region + (1|animal); see Extended Data Fig. 7a for pairwise comparisons between regions). There was a trend for isocortical regions to generate flickers at a higher rate than subcortical regions ($P = 0.071$, Spearman rank correlation). Consistent with synthetic controls (Fig. 5c), flickers were observable down to 67 ms. Mean flicker duration varied significantly by region from 142 to 416 ms (Fig. 5d, $P = 1 \times 10^{-53}$, linear mixed model: flicker duration ~ region + (1|animal); see Extended Data Fig. 7a for pairwise comparisons and Supplementary Fig. 4a,b for distributions). Subcortical regions produced significantly longer flickers than isocortical regions ($P = 0.037$, Spearman rank correlation).

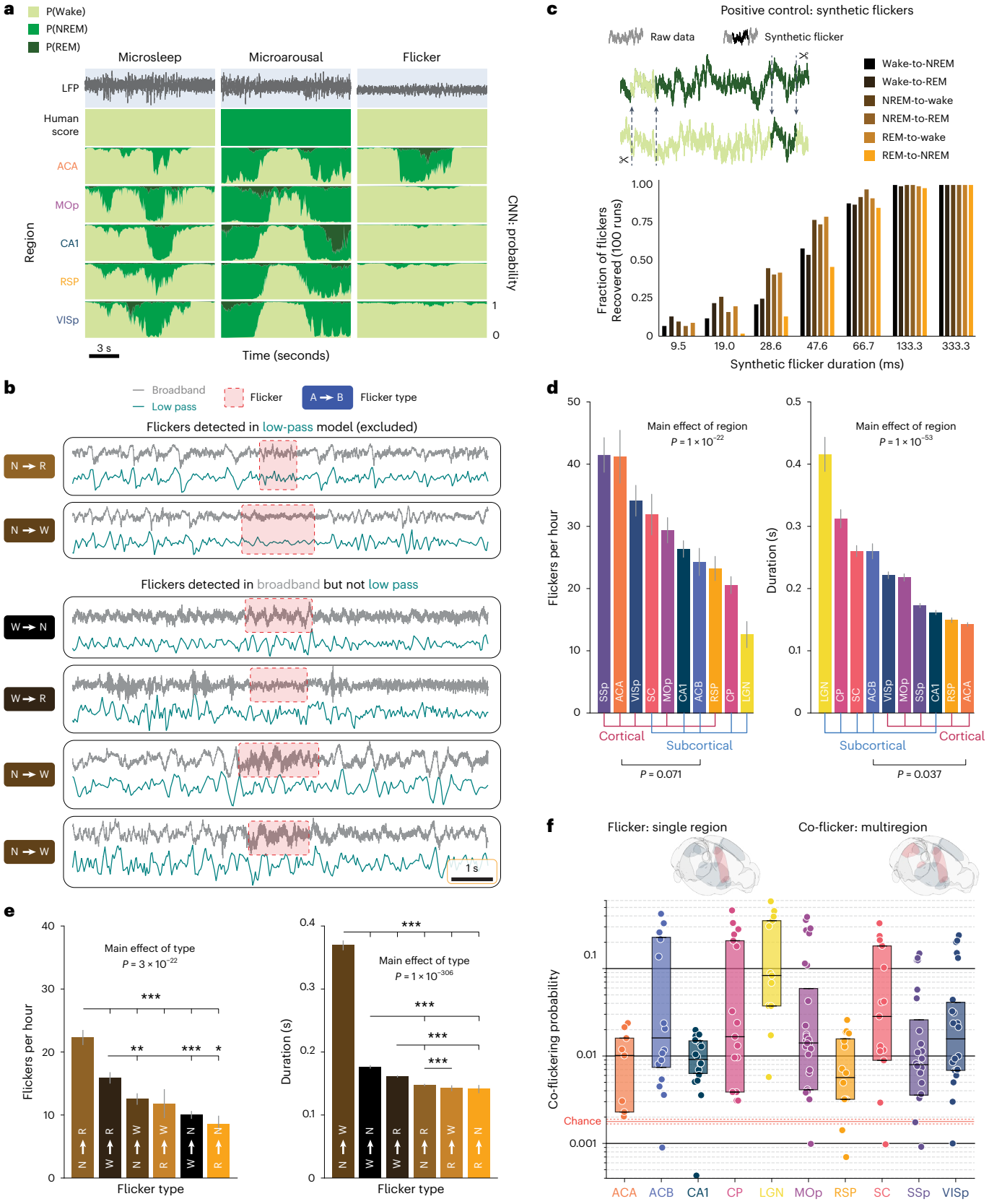
All six possible flicker types, that is, combinations of surrounding and flicker states, occurred in our data. There was a significant effect

Fig. 5 | Individual regions briefly switch states independently of the rest of the brain. a, Examples of three forms of disagreement between CNN classification and human consensus scoring of brain state. The top trace is neural broadband. The second row is human scoring of the corresponding state. The bottom four rows are outputs of independent CNNs trained in each of the four brain regions recorded in the same animal. Specifically, they show the predicted probability for each state (P(state)). The left column is a microsleep: all regions (global) show a brief, high-confidence intrusion of sleep into surrounding wake. The center is a microarousal: all regions show a brief, high-confidence intrusion of wake into surrounding sleep. The right demonstrates a wake-to-NREM 'flicker' in anterior cingulate. Flickers are defined as high-confidence, nonglobal events that are not detected in low-pass models (**b**), distinct from transitions between states. **b**, Brief local events were identified in 16 Hz low-pass models as well as broadband models. Events detected in both were excluded from analyses. Examples are shown in the top two rows: flickers are detected in both broadband (gray trace) as well as low-pass (teal). The red boxes denote the interval identified as a flicker in each model. The flicker type is shown on the left. The bottom four rows are examples of flickers identified in the broadband but not low-passed data. **c**, Top: schematic of synthetic-flicker-positive control. Short segments of data from each state were transposed into segments of every other state. Bottom: the proportion of synthetic flickers predicted by CNNs in at least one timestep as a function of duration and flicker type. **d**, Flicker rate and duration varied significantly as a function of region. There was a trend toward higher flicker rates in isocortical compared to subcortical regions ($P = 0.071$, Spearman rank correlation). Subcortical regions exhibited significantly longer flickers than isocortical regions ($P = 0.037$, Spearman rank correlation). Linear mixed models, (for example, flicker rate ~ region + (1|animal)) with ANOVA for main effect and EMMeans with Tukey post hoc for comparisons. See Extended Data Fig. 7a for individual comparisons. **e**, The mean rate (left) and duration (right) of each flicker type. Linear mixed models, for example, flicker rate ~ type + (1|animal)) with ANOVA for main effect and EMMeans with Tukey post hoc for comparisons. See Extended Data Fig. 7b for individual comparisons. n.s., $P > 0.05$. *$P \leq 0.05$, **$P < 0.01$, ***$P < 0.001$. **f**, Coincident flickering in two or more anatomically distinct regions occurred significantly above chance ($P < 0.001$, permutation test). Top: illustration of co-flickers versus individual flickers. Bottom: box plot of probability of co-flickering by region. The box plots show intermediate quartiles. The swarms, colored by region, show all pairs of regions within each animal. The shaded red area between the dashed red lines indicates the range of chance in all regions. The solid red line is the mean chance level. $n = 45$ implantation sites, 9 animals. The data are presented as mean ± s.e.m. Bonferroni correction was used for multiple comparisons.

https://doi.org/10.1038/s41593-024-01715-2

of flicker type when comparing rate (Fig. 5e, $P = 3 \times 10^{-22}$, linear mixed model: flicker rate ~ type + (1|animal); Extended Data Fig. 7b). The most common flicker was NREM to REM, while the least common was REM to NREM. Similarly, there was a significant effect of flicker type on duration (Fig. 5e, $P = 1 \times 10^{-306}$; Extended Data Fig. 7b, and see Supplementary Fig. 4c,d for distributions). NREM-to-wake flickers were the longest, while REM-to-NREM flickers were the shortest.

Interregional coordination is known to play a major role in the emergence and maintenance of sustained arousal states as well as the transient modulation of attentional states[10,11,46,47]. We hypothesized



https://doi.org/10.1038/s41593-024-01715-2

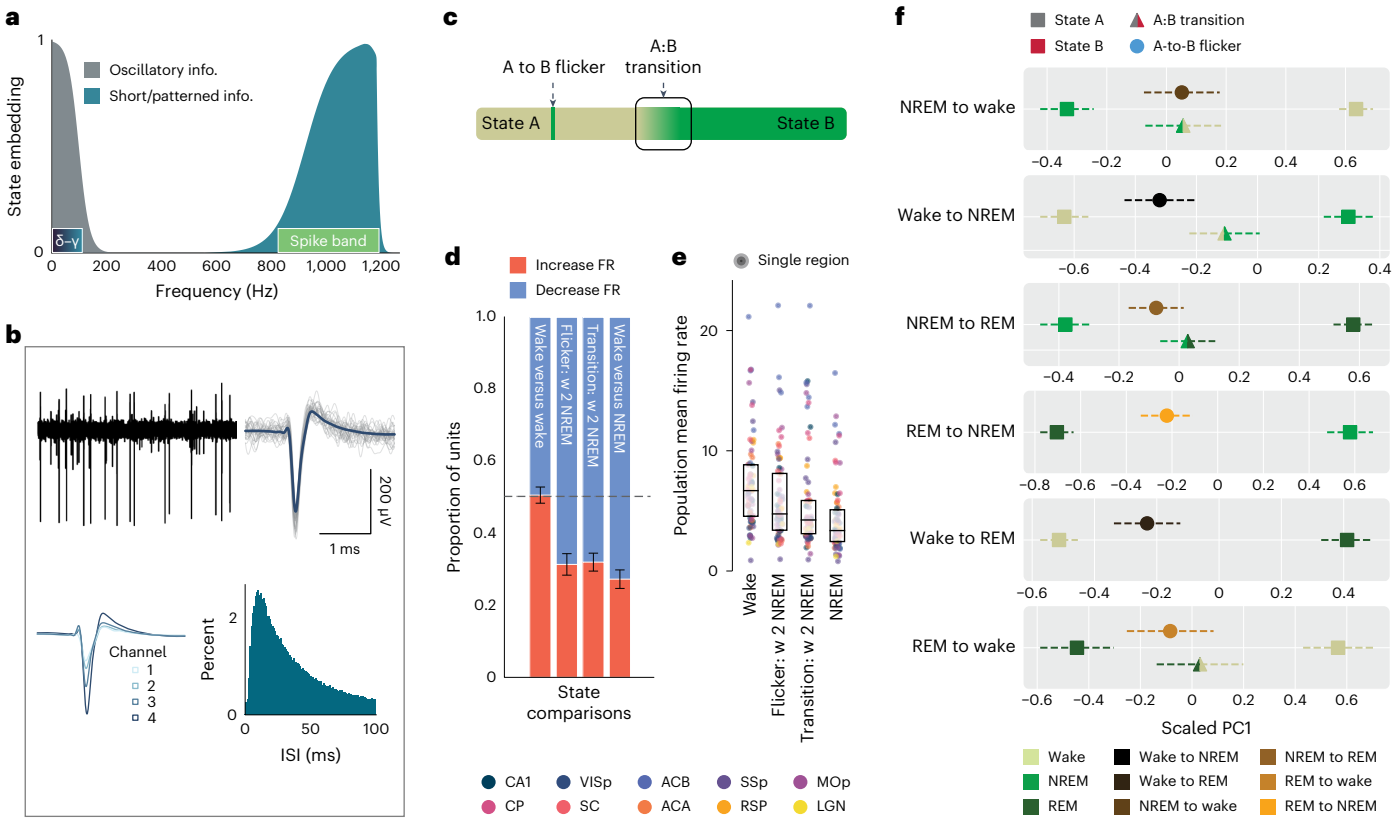


Fig. 6 | Single neuron spiking shows evidence of flickers detected by CNNs. **a**, Schematic of state information (info.) as a function of frequency. The shaded gray illustrates state information conveyed by canonical oscillations (δ–γ: bar inset). The shaded teal illustrates state embedding in the kHz range that contains action potential information (spike band). **b**, Example of a single unit. Top left: a broadband trace showing high signal-to-noise spiking. Top right: the mean waveform (dark blue) of an extracted and spike-sorted single unit (individual traces shown in gray). Bottom left: the mean waveform across the four channels of a tetrode. Bottom right: a histogram of the unit's ISIs. Note the presence of a refractory period around 0–5 ms. **c**, Conceptual illustration of a flicker and a transition. **d**, Single units may increase or decrease their firing rate as a function of state. As a control, the left bar shows that random samples of single unit firing during wake (w) are equally likely to be above or below their own waking mean. In contrast, >50% of units decrease their instantaneous rates during NREM, flickers to NREM and transitions to NREM compared to wake (right bars) (Extended Data Fig. 8a). **e**, Mean single-unit firing rate by region (color) during wake, wake-to-NREM flickers, wake-to-NREM transitions and NREM. The box plots show intermediate quartiles. The swarm shows mean rates colored by region. See Extended Data Fig. 8b. **f**, Scaled PC1 projections for the six flicker types, represented graphically by the surrounding state (state A: left square), predicted state (state B: right square), A-to-B flicker (triangle) and A-to-B transition (circle). The error bars are s.e.m. multiplied by the square root of n animals (Methods). Supplementary Tables 2–5 show significance of pairwise comparisons. $n = 45$ (**d** and **e**) and 44 (**f**) implantation sites, 9 animals. One site was excluded from PCA owing to single-unit yield.

that, if flickers represent neurobiologically meaningful substates, their timing might reveal functional connectivity between regions. In other words, flickering in one region might influence the state-related dynamics in downstream regions. Consistent with this, subnetworks of regions reliably exhibited coincident flickering far above chance (Fig. 5f). It is interesting to note that the LGN exhibited the highest rate of co-flickering, consistent with the broad connectivity of the thalamus.

Flickers were not explained by substates (off states, ripples and/or spindles), indicating that flickers are not an artifact of transient electrophysiological events (Extended Data Fig. 5c,d).

Flickering corresponds to transition-like spiking activity

Our results thus far demonstrate that CNNs identify unique state-related information in the 1 kHz range, which contains action potentials (Fig. 6a). We hypothesized that single-unit spiking might show evidence of state alterations during flickers. If so, this would suggest that flickering coincides with a shift in neuronal output. Alternately, if unit activity were unaffected by flickering, flickers detected in broadband data would probably be statistical anomalies.

We spike-sorted all raw data and extracted ensembles of well-isolated single units from every region (Fig. 6b and Supplementary Fig. 3). We separated spiking into bins corresponding to each state, transitions between states, and flickers (Fig. 6c). Many units systematically changed their firing rate as a function of state; the proportion of units whose rate varied by state was maintained in both transitions and flickers (Fig. 6d and Supplementary Fig. 10a). Likewise, population mean firing rates showed consistent shifts between states, in transitions and in flickers (Fig. 6e; all combinations shown in Supplementary Fig. 10b).

Lastly, we sought to compare the effect of flickers and transitions on individual units whose activity was sensitive to state; it is these units in particular that might be expected to be modulated by flickering. We used principal component analysis (PCA) to examine correlations between the spiking patterns of single units (sets of ten interspike intervals (ISIs)) on a region-by-region basis. By definition, the first principal component (PC1) represented the largest source of variance in unit activity; fortuitously, the axis defined by PC1 reliably separated each pair of states with near-perfect accuracy, suggesting that ensemble spiking activity is sufficient to support arousal state classification.

We then projected spiking activity during flickers and transitions onto these axes, allowing us to directly ask whether spiking during flickers showed evidence of a separation from the surrounding state. Ensemble spiking during flickers and transitions was significantly

https://doi.org/10.1038/s41593-024-01715-2

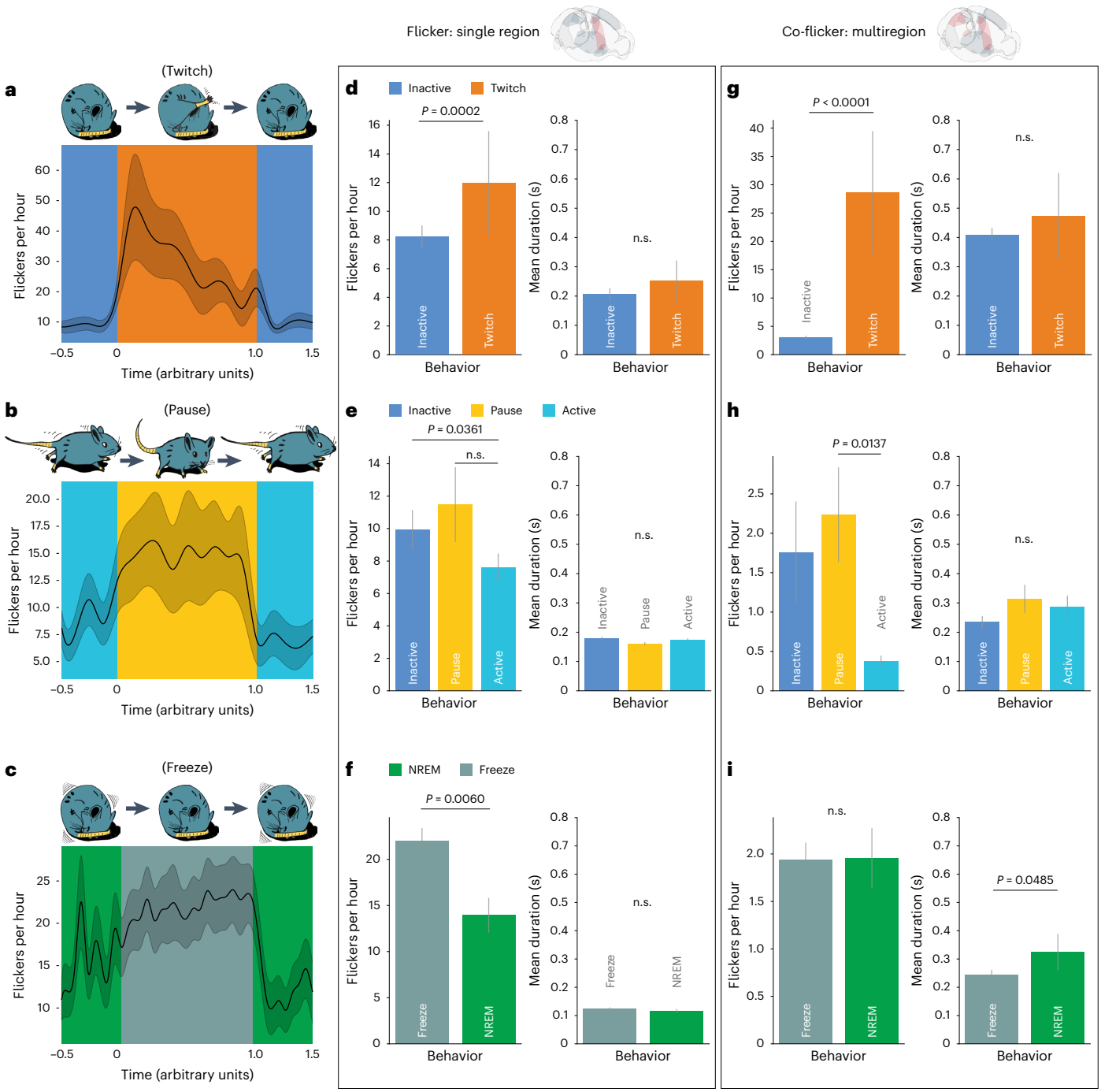


Fig. 7 | Flickering predicts structure in free behavior. a, Top: illustration of a twitch during NREM sleep. Bottom: NREM-to-wake flicker rate (both single- and multiregion flickers) before (dark blue), during (orange) and after (dark blue) twitches. Activity color codes are defined in **d**–**f**. **b**, Top: illustration of a brief pause during extended locomotion. Bottom: wake-to-NREM flicker rate (both single- and multiregion flickers) before (light blue), during (gold) and after (light blue) pauses. **c**, Top: illustration of brief 'freezing' during NREM sleep (that is, a small but significant reduction in slight movements associated with muscle tone, respiration and so on; Extended Data Fig. 1g). Bottom: NREM-to-REM flicker rate (both single- and multiregion flickers) before (green), during (slate) and after (green) freezing. **d**, Rates (left) and durations (right) of NREM-to-wake single-region flickering (top brainrender) as a function of motion states shown in **a**. **e**, Same as **d** but for wake-to-NREM flickers during the states shown in **b**. **f**, Same as **d** but for NREM-to-REM flickers during the states shown in **c**. **g**, Rates (left) and durations (right) of NREM-to-wake co-flickers (multiregion: top brainrender) as a function of the motion states shown in **a**. **h**, Same as **g** but for wake-to-NREM flickers during the states shown in **b**. **i**, Same as **g** but for NREM-to-REM flickers during the states shown in **c**. Error is s.e.m. Linear mixed effects: for example, log(flicker rate) ~ motor state × flicker type × n regions + (1 | animal/region). EMMeans and Tukey post hoc. $n = 45$ regions, 9 mice. For distributions, see Supplementary Fig. 5.

shifted from the surrounding state toward the flicker state in all six flicker types and all four transition types (Fig. 6f). Transitions were significantly separable from the surrounding state in all regions, and flickers were separable in 9/10 regions (Extended Data Fig. 8c and Supplementary Tables 2–5). In 5/6 flicker types, flicker duration was uncorrelated with position along this axis (the exception was NREM-to-wake flickers, $P < 0.001$). This suggests that even very short flickers coincide with spiking changes. These patterns could manifest in

joint population-level spike patterning (for example, spikewords) given that individual neurons generally fire at prohibitively slow rates. The number of regions a flicker was detected in was also not significantly correlated with its position along this axis. Thus, time periods identified as flickers and transitions by the CNN correspond to significant alterations in single-unit activity.

Spiking during flickers and transitions was not significantly different in 3/4 transition types and 9/10 regions (Supplementary Tables 2–5). Note that, while six flicker types were detected, only four transition types were available for comparison because global transitions from wake to REM and REM to NREM are not observed. *k*-Means clustering was applied along these axes to explore the relationship between transitions and flickers on a recording-by-recording basis. On each PC1 axis, an optimal *k* was determined by the silhouette score. We found $k = 2.79 \pm 0.10$, and flickers co-clustered more frequently with transitions (68.53%) than the surrounding (63.13%) or CNN-predicted states (34.14%). In some cases, flickers and transitions co-clustered with the CNN-predicted or surrounding states. However, despite substantial variability, this suggests the primary configuration along this axis is three clusters: one where the majority is the surrounding state, a second shared by transitions and flickers, and a third populated by the CNN-predicted state. Taken together, flickers are consistent with regional dynamics during transitions between the surrounding and CNN-predicted state.

Flickering correlates with transient behavioral changes

Our data suggest that the minimal reliably resolvable unit of brain state is on the order of 10^1 ms and arises from local neuronal activity. One possibility is that this is an epiphenomenon—in other words, resolving states at this scale provides no insight into brain function. Alternatively, understanding states at fast and local resolution could offer novel insight into the structure of natural behavior. We hypothesized that, if flickers represent momentary, regional discontinuities in state, they should coincide with similar discontinuities in behavior[48].

To detect behavioral discontinuities, we used optical flow as a coarse but sensitive measure of animal activity. Optical flow differed significantly between all pairs of states (Extended Data Fig. 1g), and the slight but significant reduction in optical flow during REM relative to NREM is consistent with REM paralysis. We algorithmically identified three forms of behavioral discontinuity: (1) motor twitches during extended NREM sleep (Fig. 7a), (2) brief pauses amid extended locomotor sequences (Fig. 7b) and (3) momentary reductions in slight movements during NREM sleep, such as those associated with muscle tone and respiration (that is, 'freezing'; Fig. 7c). In all three examples of brief behavioral switching, we observed increased flicker rates. Specifically, NREM-to-wake flickers were enriched during sleep twitches (Fig. 7a, $P < 0.001$, linear mixed model: flicker rate ~ motor state × flicker type × *n* regions + (1|animal/region)), wake-to-NREM flickers were enriched during brief pauses in high activity (Fig. 7b, $P = 0.0018$) and NREM-to-REM flickers trended toward an increase during freezing in sleep (Fig. 7c, $P = 0.0722$).

Curious to know whether a potential relationship might depend on the number of regions contributing to a flicker, we divided our analyses into single-region flickers and multiregion co-flickers. In the first two cases (NREM to wake and wake to NREM), multifold increases in co-flickering appeared to drive these effects, suggesting that, when more regions are recruited during a flicker, there is more impact on behavioral structure (Fig. 7d,e,g,h; see Supplementary Fig. 5 for distributions). Paradoxically, co-flickering masked a significant enrichment of single-region NREM-to-REM flickers during freezing in sleep ($P = 0.006$) (Fig. 7f,i; see Supplementary Fig. 5 for distributions).

Given the strong correlation of the CNN's classification with a change in movement, we sought to confirm that its predictions did not arise from noise artifacts (particularly, muscle activity). Crucially, such artifacts are spatially distributed and, thus, can be mitigated by removing the common signal from each individual channel. Therefore, we subtracted the common mean of all channels per region (64 channels) from the broadband signal of each wire (Extended Data Fig. 9a). We then tested our pretrained CNNs on this denoised version of the data. Because of the precise pattern recognition at the heart of deep-learning architectures, any manipulation of the input after training has the potential to render the model useless, even if relevant patterns still exist. Despite this likelihood, we observed that, in 66% of the models (81/122), accuracy was not compromised by removal of the common mean (Extended Data Fig. 9b). This was true for 2.6 s and 40 ms models. This demonstrates that the principle of fast state embedding is not artifactual and does not require spatially distributed information, such as EMG/EOG artifact, or even large-scale multiunit bursting that is picked up across a local region. The 34% of models impaired by common mean subtraction either (a) converged on a solution that took advantage of common signal (which does not imply an absence of a reliable local signal) or (b) learned a local signal that was disrupted by common mean subtraction.

While the general performance of the CNN, which (in combination with all analyses above) suggests an embedding of states via complex patterns that are fast and local, was not dependent on common signal, we sought to confirm that flickers were similarly independent. Patterns of disagreements between CNN and human labeling were largely unaffected by mean subtraction. Flickers, which by definition represent a high-confidence, nonstochastic subset of disagreements, were preserved (Extended Data Fig. 9c–e).

These data suggest that short neurophysiological states observed in a subset of regions correlate with short behavioral changes within extended sleep and wake bouts. The relationship between flickers and behavior implies that a definition of brain state based on dynamics in the kilohertz frequency range on a timescale of <100 ms may provide meaningful insight into brain function and organization.

Discussion

Sleep and wake states are widely assumed to arise from extended changes in global patterns of brain activity[2,3,5,6]. We report the unexpected finding that sleep and wake determine distinct, millisecond-scale patterns in small regions throughout the brain. We find that individual regions routinely switch states ('flicker') independently of other regions. Flickers demonstrate local, fast, state-related patterning of neuronal activity. Flickers, as brief neurophysiological discontinuities, are enriched in brief behavioral discontinuities in both sleep and wake. This implies that fast, state-related dynamics are implicated in cognition and behavior. Our data suggest the existence of a fundamental unit by which state determines brain activity that is distinct from low-frequency oscillations.

Our data also suggest that extremely short intervals of neural data contain latent structure that is determined by state. This takes the form of distinct temporal patterns, although the mechanism behind such a phenomenon remains unclear. One possibility is nested oscillations, similar to how 1 Hz slow waves in NREM coordinate sharp wave ripples (150 Hz). However, this is unlikely for three reasons. First, fast embedding exists despite slow rhythms, a property illustrated by flickering (for example, Fig. 5b, bottom two rows. High-frequency information is altered during flickers despite ongoing slow waves). Second, there is a 500 Hz gap between low-frequency rhythms and fast embedding (Figs. 1g,h and 6a, and Extended Data Fig. 4). Third, analysis of nested events and substates (Fig. 4) reveals that, amid dramatic changes in low-frequency activity, reliable fast embedding is trivially recoverable. As a result, it is not immediately clear how traditional wave information below 100 Hz could establish the kilohertz-frequency fast embedding described here.

However, extensive empirical evidence suggests that slow, broad signals are the foundation of sleep and wake states in the brain[2,49].

Consistent with this, the activation of broadly projecting nuclei in the midbrain and brainstem drives changes in brain state[50–53]. Our data thus appear paradoxical: sleep and wake can be experimentally controlled by slow and anatomically distributed mechanisms yet cannot account for millisecond patterns nor local switches in state. This incongruity could be solved by simply shifting the mechanism of state from the global signal to the local region. In this framework, traveling waves and neuromodulatory tone function as coordinating signals, instructing the state of distributed regions. From this perspective, the fundamental unit of sleep and wake states might comprise nonoverlapping libraries of spike patterns available to each region. Given the diversity of neuronal activity observed across the brain, we suspect that these highly localized state-dependent patterns may be unique to the region and not generalizable to others. A slow and global signal would generally coordinate regions[23,24], thus determining the behavioral macrostate of the organism. Still, important work will reveal how spatial heterogeneity in fast embedding is related to spatial variation in organizing signals, such as neuromodulatory tone[54].

Shifts in neuronal activity correlate with brain state in many species and brain regions[55–59]. Detection of such changes typically requires extended observation, and the effects are too variable for use as a classification tool[31,60]. Here, small intervals of ensemble spiking activity are highly effective in separating each state (Fig. 6f). This is due to treating each neuron independently; it appears that, across many regions, individual neurons are diverse in their state-dependent activity and more reliable than population measures. However, this is less generalizable: it requires a priori knowledge of each neuron's profile and the ability to track cells over extended periods of time.

The 'black box' nature of machine learning tools represents a key challenge to the field. While our models are empirically effective, answering the question of what they are learning in small intervals of time and space is crucial. Broadly, there are two approaches to interpretability/attribution. First, it is possible to directly query the inner workings of a model, via methods such as feature visualization[61]. This is particularly effective in models that classify images; feature visualization reveals ghostly simulacra of the training data. The second approach is progressive ablation. By carefully whittling away different features of the input data, one can reveal the fundamental source of information necessary for a model to learn. Ultimately, our models learn well given only 10^1 ms of 750 Hz high-passed data from a single channel. This cannot be explained by single-unit spike waveforms, as a unit would need to be spiking every 40 ms (that is, 25 Hz tonically) for continuous classification. As a result, visualization could only reveal a series of uninterpretable traces. Instead, we employed a thorough and rigorous ablative approach to understand the origins and constraints of the source of the underlying signal learned by the model.

Microarousals and microsleeps have received attention in both humans[62,63] and rodents[16,64]. There is disagreement regarding precise definitions that causes subtle inconsistencies in their detection between groups[65,66]. However, these are common experiences, such as momentarily slipping into sleep when drowsy. Local states have also been described previously. Utilizing sleep deprivation, Vyazovskiy et al.[17] demonstrated the ability of the waking cortex to support patches of slow wave activity, that is, 'local sleep'. This influential work led to the identification of wake-like and sleep-like oscillatory activity during sleep and wake, respectively, in humans[14,67,68] as well as rats[15]. In our data, we found robust evidence of these previously described phenomena, each of which was based on low-frequency rhythms. By excluding these events from our subsequent analyses of flickers, we demonstrate that the contribution of fast embedding to brain function is distinct from brief and local alterations in low-frequency rhythms. However, flickers between wake and NREM are behaviorally consistent in several respects with local sleep, microsleeps and microarousals. Thus, it may be appropriate to consider flickers as a point along the increasingly rich spatiotemporal continuum of states[7,45]. Collectively, increasing evidence of local manifestations of state supports a view that wake and sleep are not purely descriptors of the behavioral state of an animal.

We were initially surprised to record robust evidence of all six possible types of flicker (Fig. 5e and Extended Data Fig. 10), given that some pairings of states do not map onto transitions observed under normal conditions, in particular, wake-to-REM flickers. Despite their improbability, these features are clearly evident in our data even when defined with relatively conservative inclusion criteria. Similar transitions in the vertebrate brain are not impossible: in narcolepsy, for example, REM-like neural activity arises during wake[69,70]. Given the ability of CNNs to reliably recover synthetic wake-to-REM flickers and their presence in all animals and all regions, there is clearly a latent, high-frequency signature of REM that arises during normal waking. How this relates to the content of behavior remains unclear. NREM-to-wake and wake-to-NREM flickers connect to behavior intuitively: as more regions are involved, the behavior is more concrete. Contrastingly, NREM-to-REM flickers exhibit the opposite trend. One possibility is that future data will resolve this disparity, as NREM-to-REM flickers were relatively sparse given the scale of the rest of our data. Alternatively, should this pattern hold, it implies that, as NREM-to-REM flickers engage more regions, the behavioral impact is actively counterbalanced.

In this work, we took a neural-activity forward approach to the question, 'what is the fundamental unit of brain state'? We approached this with a first-principles perspective that the fundamental basis of sleep/wake is the smallest window in which sleep/wake determines neuronal activity. This is analogous to the molecule as the minimal unit of matter: while a given type of matter may be more easily studied and characterized at larger scales, it cannot exist below that of the molecule. To assess this empirically, we asked, 'what is the minimal resolvable signature of sleep and wake states'? We have applied high-resolution, multisite, continuous, long-term observations of neural dynamics to reach a first approximation of such a limit. This data-driven limit is almost certainly an underestimation of the true minimum. Future work with more advanced methodology will probably push these limits further. However, our observations suggest that the canonical heuristics by which sleep has been defined for nearly a century are not the fundamental basis of organization of state (as defined above). Two immediate avenues for future work hold great promise. First, it will be fascinating to approach this problem in an unsupervised fashion; the number of reliably detectable substates may be far larger than the number of states observable with traditional heuristics. Second, a behavior-forward approach to state has great potential to further connect neuronal dynamics and brain function.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-024-01715-2.

References

1. Berger, H. Über das Elektrenkephalogramm des Menschen. *Arch. Psychiatr. Nervenkrankh.* **87**, 527–570 (1929).
2. Gervasoni, D. et al. Global forebrain dynamics predict rat behavioral states and their transitions. *J. Neurosci.* **24**, 11137–11147 (2004).
3. Volgushev, M. et al. Precise long-range synchronization of activity and silence in neocortical neurons during slow-wave sleep. *J. Neurosci.* **26**, 5665–5672 (2006).
4. Burle, B. et al. Spatial and temporal resolutions of EEG: is it really black and white? A scalp current density view. *Int. J. Psychophysiol.* **97**, 210–220 (2015).
5. Ding, F. et al. Changes in the composition of brain interstitial ions control the sleepwake cycle. *Science* **352**, 550–555 (2016).

6. Lee, S.-H. & Dan, Y. Neuromodulation of brain states. *Neuron* **76**, 209–222 (2012).
7. Nir, Y. & de Lecea, L. Sleep and vigilance states: embracing spatiotemporal dynamics. *Neuron* **111**, 1998–2011 (2023).
8. Routtenberg, A. Hippocampal correlates of consummatory and observed behavior. *Physiol. Behav.* **3**, 533–535 (1968).
9. Sainsbury, R. S., Heynen, A. & Montoya, C. P. Behavioral correlates of hippocampal type 2 theta in the rat. *Physiol. Behav.* **39**, 513–519 (1987).
10. Harris, K. D. & Thiele, A. Cortical state and attention. *Nat. Rev. Neurosci.* **12**, 509–523 (2011).
11. Engel, T. A. et al. Selective modulation of cortical state during spatial attention. *Science* **354**, 1140–1144 (2016).
12. Lacroix, M. M. et al. Improved sleep scoring in mice reveals human-like stages. Preprint at *bioRxiv* https://doi.org/10.1101/489005 (2018).
13. Huber, R. et al. Arm immobilization causes cortical plastic changes and locally decreases sleep slow wave activity. *Nat. Neurosci.* **9**, 1169–1176 (2006).
14. Nir, Y. et al. Regional slow waves and spindles in human sleep. *Neuron* **70**, 153–169 (2011).
15. Emrick, J. J. et al. Different simultaneous sleep states in the hippocampus and neocortex. *Sleep* **39**, 2201–2209 (2016).
16. Soltani, S. et al. Sleep–wake cycle in young and older mice. *Front. Syst. Neurosci.* **13**, 51 (2019).
17. Vyazovskiy, V. V. et al. Local sleep in awake rats. *Nature* **472**, 443–447 (2011).
18. Rattenborg, N. C. et al. Evidence that birds sleep in mid-flight. *Nat. Commun.* **7**, 12468 (2016).
19. Serafetinides, E. A., Shurley, J. T. & Brooks, R. E. Electro-encephalogram of the pilot whale, *Globicephala scammoni*, in wakefulness and sleep: lateralization aspects. *Int. J. Psychobiol.* **2**, 129–135 (1972). [Google Scholar].
20. Tamaki, M. et al. Night watch in one brain hemisphere during sleep associated with the first-night effect in humans. *Curr. Biol.* **26**, 1190–1194 (2016).
21. Rector, D. M. et al. Local functional state differences between rat cortical columns. *Brain Res.* **1047**, 45–55 (2005).
22. Amzica, F. & Steriade, M. Electrophysiological correlates of sleep delta waves. *Electroencephalogr. Clin. Neurophysiol.* **107**, 69–83 (1998).
23. Buzsáki, G. & Schomburg, E. W. What does gamma coherence tell us about interregional neural communication? *Nat. Neurosci.* **18**, 484–489 (2015).
24. Mölle, M. et al. Hippocampal sharp wave-ripples linked to slow oscillations in rat slow-wave sleep. *J. Neurophysiol.* **96**, 62–70 (2006).
25. Girardeau, G. & Lopes-dos-Santos, V. Brain neural patterns and the memory function of sleep. *Science* **374**, 560–564 (2021).
26. Muñoz-Torres, Z. et al. Amygdala and hippocampus dialogue with neocortex during human sleep and wakefulness. *Sleep* **46**, zsac224 (2022).
27. Rolnick, D. et al. Deep learning is robust to massive label noise. Preprint at http://arxiv.org/abs/1705.10694 (2018).
28. Gent, T. C., Bassetti, C. L. A. & Adamantidis, A. R. Sleep–wake control and the thalamus. *Curr. Opin. Neurobiol.* **52**, 188–197 (2018).
29. Saper, C. B. Staying awake for dinner: hypothalamic integration of sleep, feeding, and circadian rhythms. In *Hypothalamic Integration of Energy Metabolism, Proc. 24th International Summer School of Brain Research, held at the Royal Netherlands Academy of Arts and Sciences* 243–252 (Elsevier, 2006).
30. Ellis, C. A., Miller, R. L. & Calhoun, V. D. A systematic approach for explaining time and frequency features extracted by convolutional neural networks from raw electroencephalography data. *Front. Neuroinform.* **16**, 872035 (2022).
31. Hengen, K. B. et al. Neuronal firing rate homeostasis is inhibited by sleep and promoted by wake. *Cell* **165**, 180–191 (2016).
32. Chung, J. E. et al. A fully automated approach to spike sorting. *Neuron* **95**, 1381–1394.e6 (2017).
33. Bédard, C., Kröger, H. & Destexhe, A. Model of low-pass filtering of local field potentials in brain tissue. *Phys. Rev. E* **73**, 051911 (2006).
34. Harris, K. D. et al. Improving data quality in neuronal population recordings. *Nat. Neurosci.* **19**, 1165–1174 (2016).
35. Trautmann, E. M. et al. Accurate estimation of neural population dynamics without spike sorting. *Neuron* **103**, 292–308.e4 (2019).
36. Vanderwolf, C. H. Hippocampal electrical activity and voluntary movement in the rat. *Electroencephalogr. Clin. Neurophysiol.* **26**, 407–418 (1969).
37. Girardeau, G. et al. Selective suppression of hippocampal ripples impairs spatial memory. *Nat. Neurosci.* **12**, 1222–1223 (2009).
38. Karlsson, M. P. & Frank, L. M. Awake replay of remote experiences in the hippocampus. *Nat. Neurosci.* **12**, 913–918 (2009).
39. Kay, K. et al. A hippocampal network for spatial coding during immobility and sleep. *Nature* **531**, 185–190 (2016).
40. Vallat, R. & Walker, M. P. An open-source, high-performance tool for automated sleep staging. *eLife* **10**, e70092 (2021).
41. Funk, C. M. et al. Local slow waves in superficial layers of primary cortical areas during REM sleep. *Curr. Biol.* **26**, 396–403 (2016).
42. Halasz, P. Hierarchy of micro-arousals and the microstructure of sleep. *Neurophysiol. Clin.* **28**, 461–475 (1998).
43. Ekstedt, M., Åkerstedt, T. & Söderström, M. Microarousals during sleep are associated with increased levels of lipids, cortisol, and blood pressure. *Psychosom. Med.* **66**, 925–931 (2004).
44. Andrillon, T. et al. Predicting lapses of attention with sleep-like slow waves. *Nat. Commun.* **12**, 64–78. (2021).
45. Siclari, F. & Tononi, G. Local aspects of sleep and wakefulness. *Curr. Opin. Neurobiol.* **44**, 222–227 (2017).
46. Poulet, J. F. A. & Petersen, C. C. H. Internal brain state regulates membrane potential synchrony in barrel cortex of behaving mice. *Nature* **454**, 881–885 (2008).
47. Tan, A. Y. Y. et al. Sensory stimulation shifts visual cortex from synchronous to asynchronous states. *Nature* **509**, 226–229 (2014).
48. Kramer, D. L. & McLaughlin, R. L. The behavioral ecology of intermittent locomotion. *Am. Zool.* **41**, 137–153 (2001).
49. Steriade, M., McCormick, D. A. & Sejnowski, T. J. Thalamocortical oscillations in the sleeping and aroused brain. *Science* **262**, 679–685 (1993).
50. Carter, M. E. et al. Tuning arousal with optogenetic modulation of locus coeruleus neurons. *Nat. Neurosci.* **13**, 1526–1533 (2010).
51. Chen, K.-S. et al. A hypothalamic switch for REM and non-REM sleep. *Neuron* **97**, 1168–1176.e4 (2018).
52. Moruzzi, G. & Magoun, H. W. Brain stem reticular formation and activation of the EEG. *Electroencephalogr. Clin. Neurophysiol.* **1**, 455–473 (1949).
53. Li, S.-B. et al. Hyperexcitable arousal circuits drive sleep instability during aging. *Science* **375**, eabh3021 (2022).
54. Sweyta Lohani et al. Spatiotemporally heterogeneous coordination of cholinergic and neocortical activity. *Nat. Neurosci.* **25**, 1706–1713 (2022).
55. Noda, H. & Adey, W. R. Changes in neuronal activity in association cortex of the cat in relation to sleep and wakefulness. *Brain Res.* **19**, 263–275 (1970).
56. Abásolo, D. et al. Lempel–Ziv complexity of cortical activity during sleep and waking in rats. *J. Neurophysiol.* **113**, 2742–2752 (2015).
57. Watson, B. O. et al. Network homeostasis and state dynamics of neocortical sleep. *Neuron* **90**, 839–852 (2016).

58. Levenstein, D. et al. Sleep regulation of the distribution of cortical firing rates. *Curr. Opin. Neurobiol.* **44**, 34–42 (2017).
59. Brunwasser, S. J. et al. Circuit-specific selective vulnerability in the DMN persists in the face of widespread amyloid burden. Preprint at *bioRxiv* https://doi.org/10.1101/2022.11.14.516510 (2022).
60. Xu, Y. et al. Sleep restores an optimal computational regime in cortical networks. *Nat. Neurosci.* **27**, 1–11 (2024).
61. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. Preprint at http://arxiv.org/abs/1312.6034 (2013).
62. Torsvall, L. & Åkerstedt, T. Sleepiness on the job: continuously measured EEG changes in train drivers. *Electroencephalogr. Clin. Neurophysiol.* **66**, 502–511 (1987).
63. Carskadon, M.A. & Rechtschaffen, A. in *Principles and Practice of Sleep Medicine* (eds. Kryger, M. H., Roth, T. & Dement, W. C.) 1359–1377 (Elsevier, 2005).
64. Franken, P., Malafosse, A. & Tafti, M. Genetic variation in EEG activity during sleep in inbred mice. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **275**, R1127–R1137. (1998).
65. Kjaerby, C. et al. Memory-enhancing properties of sleep depend on the oscillatory amplitude of norepinephrine. *Nat. Neurosci.* **25**, 1059–1070 (2022).
66. Hertig-Godeschalk, A. et al. Microsleep episodes in the borderland between wakefulness and sleep. *Sleep* **43**, zsz163 (2019).
67. Nobili, L. et al. Dissociated wake-like and sleep-like electro-cortical activity during sleep. *NeuroImage* **58**, 612–619 (2011).
68. Hung, C.-S. et al. Local experience-dependent changes in the wake EEG after prolonged wakefulness. *Sleep* **36**, 59–72 (2013).
69. Kroeger, D. & de Lecea, L. The hypocretins and their role in narcolepsy. *CNS Neurol. Disord. Drug Targets* **8**, 271–280 (2009).
70. Cao, M. T. & Guilleminault, C. in *Principles and Practice of Sleep Medicine* (eds. Kryger, M. H., Roth, T. & Dement, W. C.) 873–882.e5 (Elsevier, 2017).
71. Claudi, F. et al. Visualizing anatomically registered data with brainrender. *eLife* **10**, e65751 (2021).

Methods

Mice

All procedures involving mice were performed in accordance with protocols approved by the Washington University in Saint Louis Institutional Animal Care and Use Committee, following guidelines described in the US National Institutes of Health Guide for the Care and Use of Laboratory Animals. C57BL/6 mice from Charles River were used. Seven female and two male mice were used in this study. No statistical methods were used to predetermine sample sizes, but our sample size of nine mice is similar to those reported in previous publications[17,56,60]. Mice were at least 100 days old at the beginning of recording (mean age 220 days). Mice were housed in an enriched environment and kept on a 12:12 h light:dark cycle. Ambient temperature was maintained between 68 °F and 79 °F. Humidity was maintained between 30% and 70%. Mice had ad libitum access to food and water.

Surgery

All mice underwent multisite electrode array implantation surgery. Mice were anesthetized with isoflurane (1–2% in air) and administered slow-release buprenorphine (ZooPharm, 0.1 mg kg^{-1}). The mouse's skull was secured in a robotic stereotaxic instrument (NeuroStar), and the skin and periosteum covering the dorsal surface of the skull was removed. Pitch, yaw and roll were calculated to maximize alignment with stereotaxic atlases. Three to eight craniotomies (diameter 1–1.5 mm) were drilled using the automatic drilling function of the stereotaxic robot, and the dura was resected. In each animal, each of three to eight brain regions was implanted with a custom 64-channel tetrode-based array. Arrays were fixed (not drivable) and separated from the headstage hardware by a flex cable, thus allowing an arbitrary geometry of multiple probes. Across 9 mice, a total of 45 implants spanned 10 unique brain regions. Coordinates were as follows (AP/ML/DV relative to bregma and dura, in mm): CA1 (n = 6, −2.54/−1.75/−1.5), VISp (n = 5, −3.8/−2.8/−0.8), ACB (n = 2, 1.25/−0.81/−3.94), SSp (n = 6, −0.5/2.25/−0.8), MOp (n = 7, 1.25/1.75/−1), CP (n = 4, 0.5/1.52/−2.56), SC (n = 3, −4.0/−1.0/−1.0), ACA (n = 3, 0.6/0.8/−0.94), RSP (n = 4, −1.5/0.3/−0.9) and LGN (n = 2, −2.25/2.26/−2.40). Electrode bundles were lowered into brain tissue at a rate of 5 mm min^{-1} using a custom-built stereotaxic vacuum holder. Anatomical location was confirmed post hoc via histological reconstruction (Supplementary Figs. 1b and 2). Arrays were secured with dental cement (C&B-Metabond Quick! Luting Cement, Parkell Products Inc; Flow-It ALC Flowable Dental Composite, Pentron), and headstage electronics (eCube HS-640, White Matter LLC) were bundled and secured in a three-dimensionally printed housing. Eight-module (512-channel) implants including arrays, cement and headstage electronics weighed approximately 4 g. Mice were administered meloxicam (Pivetal, 5 mg kg^{-1} per day for 3 days) and dexamethasone (0.5 mg kg^{-1} per day for 3 days) and allowed to recover in the recording chamber for at least 1 week before recording.

Recording

Recordings were made using tetrode-based arrays (12-µm-diameter gold-plated NiCr wire, Alleima LLC). Sixteen tetrodes (64 channels) were soldered to a custom-designed printed circuit board (5 mm × 5 mm × 200 µm) that plugged into an eCube amplifier module (10 × 10 × 2 mm footprint). Printed circuit board–amplifier pairs were stacked vertically with up to seven additional pairs (total eight modules, 512 channels). Recordings were conducted in an enriched home cage environment with social access to a litter mate through a perforated acrylic divider. Freely behaving mice were attached to a thin, light, highly flexible cable with in-line commutation. Neuronal signals were buffered, filtered (~0.1 Hz to 7 kHz), amplified, digitized (14 bit) and sampled at 25 kHz using the eCube Server electrophysiology system (White Matter LLC). Recordings were made continuously for between 2 weeks and 3 months (Servernode software, White Matter LLC). Synchronized 15–30 fps video was recorded concurrently (e3vision, White Matter LLC). Data were also monitored remotely using Open Ephys[72]. Twenty-four-hour blocks of data were identified for inclusion in these studies first by the absence of hardware problems, for example, cable disconnects, and second by the maximal yield of active channels. Beyond these criteria, the selection of 24 h blocks was arbitrary. The same 24 h block was utilized for all recorded circuits in an individual animal.

For experiments involving spike-sorted data, raw data were band-pass filtered between 350 and 7,500 Hz and spike waveforms were extracted and clustered using a modified version of SpikeInterface[73] and MountainSort4[32] with curation turned off. A custom XGBoost was used to identify those clusters constituting single units. Clusters identified as single units were manually inspected to confirm the presence of high-amplitude spiking, stable spike amplitude over time, consistent waveform shapes and little to no refractory period contamination.

Probe localization

Following recording, mice were perfused with 4% paraformaldehyde, and the brain was extracted and immersion fixed for 24 h at 4 °C in paraformaldehyde. The brains were then transferred to a 30% sucrose solution in phosphate-buffered saline and stored at 4 °C until they sank. Brains were then sectioned at 50 µm on a cryostat. Sections were rinsed in phosphate-buffered saline before mounting on charged slides (SuperFrost Plus, Fisher) and stained with cresyl violet. Stained sections were aligned with the Allen Institute Mouse Brain Atlas[74], and tetrode tracks were identified under a microscope (Supplementary Fig. 2b).

Consensus sleep scoring

Three experts independently scored the arousal state of each mouse using custom software (Extended Data Fig. 1d). Briefly, the local field potential (LFP) spectral power (0.1–60 Hz) was extracted from five channels selected from cortical implantations and averaged. Movement data were extracted from video recordings using DeepLabCut[75] or EMG. LFP and movement data were preliminarily sleep scored in 4 s epochs by a random forest. Human experts then evaluated LFP spectral density, movement and random forest output in 4 s epochs. Scoring software provided immediate access to temporally aligned video for disambiguation.

The three independently generated arrays of state assignments were then compared, and all disagreements were identified. The three contributing human experts then, as a group, reevaluated each epoch of disagreement and generated a consensus state label.

CNN construction and experiments

In this work, we used a single neural network model. We chose this architecture as it is particularly robust to label error and is thus well suited to tasks in which there may be substantial variation in labels[27]. Specifically, we coded a one-dimensional (1D) eight-layer fully CNN composed of seven convolutional and one fully connected layer (size 150) in TensorFlow (Python). We used a stride of four to reduce input size fourfold in each layer. This size reduction dictated the number of layers in the network to support our largest model of 65,536 inputs (2.6 s). We used a kernel size of 30, which indicates how many data points each layer sees in the convolution step. Features are built on the basis of the kernel size at each layer. Layers contained the following number of filters: layer 1 had 320, layer 2 had 384, layer 3 had 448 and layers 4–7 had 512. The output of the model is three values, a probability distribution over wake, NREM and REM. This model architecture was the same for all input sizes from one sample to 2.6 s (65,536 data points). L2 weight regularization with 1×10^{-6} was utilized. Learning rates were progressively reduced from 1×10^{-4} to 5×10^{-6} throughout training. Each layer utilized a standard ReLU activation function with the exception of the final convolutional layer, which had no activation function applied. The loss function minimized by the CNN was softmax cross-entropy[76]. CNN confidence was quantified as the entropy of the

output distribution scaled to [0, 1]. Model parameters were chosen through a manual process of hyperparameter tuning. The manual hyperparameter tuning followed a coordinate descent approach in which each hyperparameter is varied until an optimal value is identified for that parameter, then the next parameter is varied, holding all others fixed. Hyperparameter selection was conducted early in the experimentation cycle and remained fixed throughout the experiments so as not to introduce added confounds.

In all experiments, the CNN was presented with raw neural data. Two exceptions exist: (1) if data were low-passed below 16 Hz to examine the role of canonical oscillations or (2) if data were high-passed above 750 Hz to examine the role of high-frequency (predominantly neuronal) activity. When these exceptions occur, it is for the purpose of interpretation and is directly specified in text, figure legends, and often the figure itself. Models were tasked with learning to predict sleep and wake states (REM, NREM and wake) on the basis of labels generated by a consensus among three expert human scorers. In some experiments, raw neural data were progressively ablated (see below) in preprocessing. To avoid common numerical problems, input data were linearly scaled by 10^{-3}. The batch size during training was one, and the model was trained for 175,000 training steps. We selected a train–test split such that the training was composed of 18 consecutive hours of data, and the held-out test set comprised 6 h spanning a light–dark transition. Unless otherwise noted, step size was 1/15th of a second to match video frames per second. Model performance was evaluated using balanced accuracy[77,78] unless otherwise noted below. The model code will be available at http://github.com/hengenlab at the time of publication.

CNN models functioned as follows. At each time step, a model was presented with an interval of data (the duration of this interval was experimentally varied; see below). The model's task was to assign a probability of REM, NREM and wake to the central sample point. As a result, the task of each CNN was to learn the instantaneous state label at the center of a window of data, with only the information surrounding the center. Crucially, CNNs have no memory, and thus, each observed segment of data is an independent decision.

A simpler fully convolutional model was chosen over a more complex model such as ResNet with the following rationale: (1) to reduce computational burden, enabling the training of many thousands of models with varying input conditions, (2) it is not clear that more complex models benchmarked against unrelated datasets would be a better fit for this data, and (3) the existence of substantial label error dictates that there is more value in focusing on data than on model tuning.

Basic accuracy of sleep and wake states by brain region (Fig. 1d–f). To address the question of whether sleep and wake states are robustly embedded in the dynamics of the ten brain regions sampled in our recordings, we asked if CNNs could learn to decode sleep and wake from the broadband neural data within a single brain region (64 channels of data). Model input was 2.6 s (65,536 sample points) of 64 channels of raw neural data from each implant site (n = 45 implants from N = 9 animals, 45 total models).

Extended accuracy evaluation in high gamma range (Fig. 1g,h). To examine the decodability of state information within the high gamma range, we implemented per-brain region models on neural data from two distinct animals (n = 2 animals, with 12 probes in total derived from animals 2 and 7). The data had been manipulated via band-pass filtering to confine it to progressively amplified frequency information. Model input consisted of data that underwent the following band-pass filters: broadband, low-pass 16 Hz, 100–200 Hz band pass, 200–300 Hz band pass, 300–400 Hz band pass and 400–500 Hz band pass. Six models were executed per probe (72 total models). Note that EMG signal can be extracted from 300–500 Hz (ref. 57). The failure of models to learn in these bands suggests that the signal is insufficient to support state classification in the recordings. We further confirmed that, in a true EMG recording (750 Hz high-passed), models failed to train. The impact on the models' capacity to learn from these frequency-specific datasets was closely monitored. The accuracy was evaluated using balanced accuracy directly on CNN test output.

Progressive high pass (Fig. 2a). To test whether sleep and wake states could be extracted from neural data absent the canonical waves that human scorers rely on, we progressively eliminated slow components of neural dynamics from all implants within an animal and measured the impact on the models' ability to learn. Model input was ~1 s (24,576 sample points) of high-pass filtered neural data (third-order Butterworth) from all channels within an animal (n = 1 animal). Twenty-four models were run. Specifically, models were trained and tested on neural data after the following high-pass filters were applied: 0 (raw), 0.5 Hz, 1 Hz, 2 Hz, 4 Hz, 6 Hz, 8 Hz, 12 Hz, 20 Hz, 50 Hz, 100 Hz, 250 Hz, 500 Hz, 750 Hz, 1,000 Hz, 1,500 Hz, 1,600 Hz, 1,700 Hz, 1,900 Hz, 2,000 Hz, 3,000 Hz, 5,000 Hz, 7,000 Hz and 10,000 Hz. The accuracy was evaluated using balanced accuracy directly on CNN test output.

Exploration of interregional differences in low-pass, high-pass and broadband data (Fig. 2b). To assess the potential disparities between brain regions apparent in low-pass, high-pass and unaltered broadband variations of the data, we implemented an extensive suite of models corresponding to each probe across all subjects (n = 9 mice incorporating 45 probes in total). The identical model structure was applied in all scenarios, wherein the unmodified broadband data was subjected to preprocessing via a low-pass filter at 16 Hz or a high-pass filter at 750 Hz or was left unfiltered. For each model, a single probe targeting a specific brain region (comprising 64 channels) was utilized, with a data input duration of 2.6 s (encompassing 65,536 data points). The accuracy was evaluated using balanced accuracy directly on CNN test output.

Incremental diminution of input size (Fig. 2c,d). In pursuit of ascertaining the minimal temporal duration required to accurately decode sleep and wake states, we orchestrated an expansive set of models corresponding to each probe in all subjects (n = 9 mice comprising 45 probes in total) and trained them on an input size that was systematically reduced. Each model was informed by data from a single implant located within a particular brain region. A model with identical dimensions and hyperparameters was employed in each training phase, with the variation of increasingly truncated input sizes. The model's task was to generate predictions solely based on the temporal window of data presented, devoid of any prior knowledge pertaining to the animal's state. The progressive series of input sizes utilized were as follows: 2.6 s (65,536 data points), 1.3 s (32,768 data points), 327 ms (8,192 data points), 82 ms (2,048 data points), 41 ms (1,024 data points), 10 ms (256 data points), 5 ms (128 data points), 2.5 ms (64 data points), 1.3 ms (32 data points), 0.6 ms (16 data points), 0.3 ms (8 data points), 0.16 ms (4 data points), 0.08 ms (2 data points) and 0.04 ms (1 data point). The accuracy was evaluated using balanced accuracy directly on CNN test output.

We excluded models that failed to train (with ~33.3% accuracy on test set) in part because their inclusion led to nonparametric distributions, which are challenging to statistically evaluate. Their inclusion or exclusion from statistical tests had little impact on the results of these tests.

Model training and testing on low-pass filtered data (Fig. 3d). Data for this experiment were prepared using two manually selected channels, both spiking and nonspiking, with clear local field potentials. In some cases, only one channel was included due to simple computational reasons. No exclusion criteria were applied beyond selecting only high-quality channels. Each channel's data underwent two transformations: in the first condition, they were left in their original, temporally

intact form; in the second, the data were shuffled, reordering the data within each sampled segment. The shuffling process was executed by randomly drawing samples from the dataset and shuffling the data points within the sample, maintaining the same temporally intact sampling method as used for the unshuffled data.

The models were then trained on the transformed data using a variety of input sizes from 2.6 s (65,536 data points) to 0.04 ms (1 data point), generating a total of 2,028 single channel models. The model performance was evaluated using balanced accuracy for each scenario.

Model training and testing on high-pass filtered data (Fig. 3e). The same channels and range of input sizes as in Fig. 3d were used for this part of the experiment. However, the preprocessing stage was different, employing a high-pass filter set at 750 Hz to remove low-frequency details from the raw data. The models were then trained on both the unaltered and shuffled high-pass data.

Evaluation of model accuracy versus input size (Fig. 3f). The data displayed in Fig. 3f are the same data as from Fig. 3d,e, shown as a function of the input size.

Frequency analysis (Fig. 2e and Extended Data Fig. 4)

To understand the role canonical frequency plays at the 750 Hz+ range, as well as further eliminate EMG (motion artifacts) as a confounding source of information, we used a simple one-versus-all LR model trained on bandpower features in the following ranges: 1–16 Hz, 100–200 Hz, 200–300 Hz, 300–400 Hz, 400–500 Hz (EMG/motor artifact range) and 750–3,000 Hz (novel brain-state information range). In the 1–16 Hz range, bandpower bins of 4 s and 4 s windows of data were used for the LR features, in the 100–500 Hz, and 750–3,000 Hz ranges, bins of 40 ms were used over 40 ms windows of data. All LR models were evaluated using balanced accuracy using the same data as the CNN. LR models were trained on 70% of the data and evaluated on the latter, contiguous, 30% of the data. Compared LR and CNN models used single channels of data. We show the distribution of balanced accuracies across models as boxplots in Fig. 2e.

In Extended Data Fig. 4a, we compare the ratio of LR balanced accuracy to the CNN's balanced accuracy with the formula (LR balanced accuracy − 0.333)/(CNN balanced accuracy − 0.333), where 0.333 is the chance for a three-class model. This presents the ratio of accuracy relative to the CNN; a value of 1 indicates that the LR and CNN performed on par with each other. We use the 1–16 Hz range as our positive control, observing that, although the CNN outperforms the LR model, more information is contained in the low-frequency than in the high-frequency domain. In the high-frequency range, above 750 Hz, we see that some state information is discernable at the frequency spectrum; in particular, wake versus sleep can be extracted (Extended Data Fig. 4d), but the frequency domain using bandpower/LR accounts for substantially less information than the high-passed data the CNN is trained on.

Extended Data Fig. 4b provides a unity line plot comparing the balanced accuracy of the CNN (*x*-axis) input against the corresponding LR model (750–3,000 Hz). The CNN outperforms the frequency-based LR models consistently.

Extended Data Fig. 4c shows the 750 Hz+ confusion matrices for an exemplary channel with the CNN (top) and LR (bottom). Note that the CNN exhibits balanced accuracy across all three states, whereas the LR collapses to a mostly two-state solution (rarely predicting REM).

In Extended Data Fig. 4d,e, we plot the LR weights for each state (wake, NREM and REM). In Extended Data Fig. 4d, we selected the best-performing 14 of 71 LR models trained on 1–16 Hz bandpower features and averaged the weights. We subselected well-performing models to illustrate the canonical signatures of state information identified by the LR model, further validating existing understanding of sleep state dynamics. Extended Data Fig. 4e provides two samples of LR weight distributions for 750–3,000 Hz models, to illustrate that patterned signatures observed in the 1–16 Hz models do not show up. We do not belabor the point by printing averaged versions of weights in the 750–3,000 Hz range but did validate that we saw no discernible patterns. All LR models were trained using a one-versus-all classification strategy for the best representation of model weights.

Substate analysis (Fig. 4, Extended Data Fig. 5)

We detected cortical off states in a manner similar to the one used in the original publication of their discovery and connection with behavior[79]. We identified off states as time periods when no neuronal spiking was observed across all 64 channels of selected probes implanted in various cortical regions. High-amplitude noise events were excluded by using spike-sorted data rather than solely voltage threshold crossings. The minimum time length of an off state considered in our analyses was 67 ms, consistent with the step size of our CNNs used to classify state. Correct detection was confirmed with manual inspection of the broadband and spike-sorted data for a sample of detected events. We show that the timing of off states is largely uncorrelated with the timing of flickers even in flickers detected using models trained on small samples of data (40 ms) (Extended Data Fig. 5d). The exception is flickers to wake, which are negatively cross-correlated with off states (that is, flickers to wake amid sleep do not occur during cortical off states).

We detected sleep spindles using the spindles_detect function in Yet Another Spindle Algorithm (YASA)[40]. Function documentation is available at https://raphaelvallat.com/yasa/build/html/generated/yasa.spindles_detect.html. We employed the function on broadband data from all 64 channels of selected probes implanted in various cortical regions. The data were first downsampled to 1.5 kHz and then band-passed between 12 and 16 Hz. The detection threshold was set with {'rel pow': 0.2, 'corr': None, 'rms': 1.5}. YASA's IsolationForest approach to removing outliers was employed. All other hyperparameters were default as described in the function documentation. Correct detection of spindle events and not noise was confirmed with manual inspection of the broadband and processed data for a sample of detected events.

Similarly, we detected ripples using the Kay ripple detector function in ripple detection from the Eden-Kramer-Lab GitHub[39]. Package documentation is available at https://github.com/Eden-Kramer-Lab/ripple_detection. We employed the function on broadband data from all 64 channels of selected probes implanted in CA1 hippocampus. The data was first downsampled to 1.5 kHz and then band-passed between 150 and 250 Hz. Consistent with prior work in our lab[59], only ripples with amplitude >50 µV were included in downstream analyses, and ripples less than 100 ms apart were combined. Correct detection of spindle events was confirmed with manual inspection of the broadband and processed data for a sample of detected events. Further, a sample of the data was checked to confirm that spindle and ripple timing was highly cross-correlated (Extended Data Fig. 5c), consistent with prior literature[17].

Periods of active wake and quiet wake were initially identified as extended periods of high or low positional change, tracked by DeepLabCut[8]. Correct detection was confirmed by manual inspection of the video.

Flicker experiments

Flicker definition (Extended Data Fig. 6). We applied a series of criteria to CNN output to identify flickers for inclusion in our analyses. To start, we trained three 1 s CNNs to identify NREM, REM and wake on the basis of raw broadband data (1 s input) from all 64 channels contained in each implantation site ($n = 45$ implantation sites, 9 animals, 126 models). As a result, we had triplicate CNN-generated state scores for each recorded brain region. This was to avoid sporadic random error due to subtle inconsistencies in training. We used 1 s CNNs because they strike a balance between sensitivity (being able to catch true flickers)

and specificity (not predicting false flickers). Often, we observed that CNN output would preemptively begin to increase confidence in an incoming state a few seconds before a global transition. To avoid transition-related ambiguity, we generally did not consider flickers within 30 s of a global transition, but to account for the fact that some transitions were modified in time (for example, a slow transition from quiet wake to NREM sleep), we manually evaluated CNN confidence surrounding each global state transition in the 24 h of data from each implantation site in each animal in each replicate of the CNN ($n = 45$ implantations from 9 animals, 3 replicates). We extended the window of exclusion around transitions in which evidence of the incoming state was present beyond the 30 s window. We then applied a series of confidence filters to the remaining data in each replicate. To avoid general periods of low-confidence output, we identified and excluded any 35 s epochs with a mean confidence <75%. To restrict our analyses to high-confidence flickering, we next eliminated 1 s epochs with a mean confidence <75%. The general conclusions of our analyses were not reliant on a specific confidence threshold. Together, these criteria excluded 20.91% of our recordings. In each replicate, we then assigned the high-confidence state label to each time step (1/15th of a second), and collapsed the three model outputs into a single array by selecting the majority state at each interval. We then slid a 35 s rolling mode filter across the majority state array to create a label corresponding to the stable macrostate surrounding every point in time. In other words, a running mode was used to return the most common state label over a 35 s window that moved in 1/15th steps. This was ascribed as the overall stable state at the center of the window (ignoring minor fluctuations in state). We defined flickers as disagreements between these two arrays.

We had two further exclusion criteria to avoid overlap between flickers and previously described episodic arousal phenomena. First, flickers that co-occurred across all probes in the animals were excluded. Specifically, this includes microstates, such as microsleeps and microarousals. With this conservative approach, we may potentially exclude some events that are common to all recording sites but not in some unrecorded sites (that is, not truly global). Due to the 35 s modal filter of the majority state array, these had a maximum duration of ~20 s. Second, please note that flickers were detected in raw broadband data. However, to avoid overlap with events that could be visible in low-frequency data, we also detected flickers in models trained on data low-passed below 16 Hz. We excluded any flickers detected in the broadband that had any temporal overlap with a flicker detected in the low pass. This aims to exclude many of the low-frequency phenomena previously described in EEG and LFP such as local slow waves in wake and REM.

Summarily, here are the flicker criteria (in order of implementation):

1. CNN prediction (1 s) disagrees with human-scored labels. This initially helps to prevent confusion with microarousal/microsleep and see other criteria.
2. Disagreement is observed across at least two of three independently trained CNNs. This ensures it does not arise stochastically from training.
3. Disagreement is not near a human-labeled state transition. This ensures it is not part of a natural, gradual transition between two states.
4. Predictions for the surrounding interval (35 s) are confident (mean >75%). This ensures it does not arise from a generally noisy period of recording.
5. Predictions for immediate interval (1 s) are confident (mean >75%). This ensures it is a highly confident error and not simple uncertainty.
6. Disagreement is not observed in every probe in the animal (at least one agrees with the human label). This discriminates it from global events.
7. Disagreement is not observed in a model trained on low-passed data. This discriminates it from low-frequency oscillatory events.

Please note that our estimates of the duration and frequency of flickers are limited by the sensitivity of our model. It is possible that a more sensitive model might detect flickers more frequently and/or flickers with a longer time course.

We also observed high-confidence, localized CNN errors (similar to flickers) amid transitions between states (Fig. 6c). Consistent with progression along a transition's time course, we found their rate of occurrence was inversely correlated with time to/from a human-labeled state change ($P = 0.016892$). Because they have a similar confidence profile in the CNN's predictions to flickers and they reflect transition dynamics, we used these as samples of transition spiking activity. To extract these events, we performed a comparable procedure to the previously mentioned flicker detection with two exceptions: (1) rather than excluding intervals of transitions, we excluded the complement (all nontransition intervals), (2) we did not exclude general periods of low-confidence (based on 35 s window) because the time course of a transition is often a low-confidence period.

Synthetic flickers (Fig. 5c). To quantify the ability of the CNN to accurately detect brief intervals of a state B embedded within a containing state A, we constructed synthetic flickers. To do this, we identified all intervals of state A (NREM, REM and wake) that were assigned the same label by all three human scorers, as well as the standard 2.6 s CNN model (confidence >90%; see 'Basic accuracy of sleep and wake states by brain region (Fig. 1d–f)' section). High-confidence intervals were a minimum of 3 s in length. To simulate state B, we spliced segments of each state A into each other state (six total combinations of REM, NREM and wake into one another). Splices were 9.5, 19.0, 28.6, 47.6, 66.7, 133.3 and 333.3 ms. One-hundred splices of each duration were randomly selected from all high-confidence intervals in the 24 h period and pasted into a randomly selected high-confidence interval of another state. To illustrate with a specific example, consider a continuous segment of 47.6 ms of high-confidence REM. This was randomly selected from thousands of >3 s intervals. A high-confidence segment of wake was chosen at random from thousands of examples, and the 47.6 ms REM splice was pasted into a random location in the selected segment of wake. The insertion of the REM splice overwrote the corresponding section of wake data. We then asked whether splices were correctly identified in at least one timestep by a CNN with a 655 ms window and a step size of 9.5 ms. We chose this step size to establish a functional lower limit of sensitivity while maintaining a reasonable computational load. We evaluated whether the CNN correctly identified any portion of the spliced state B.

Co-flicker analysis (Fig. 5f). To understand whether flickers are a locally regulated (independent) or global phenomenon, we calculated the conditional probability of all pairs of regions flickering simultaneously according to

$$P(A|B) = \frac{P(AB)}{P(B)}.$$

Single-unit analysis (Fig. 6d–f). For each sustained arousal state (wake, NREM and REM) and flicker/transition type state (wake to NREM, wake to REM, NREM to wake, NREM to REM, REM to wake and REM to NREM), we aggregated all ISIs for each neuron present during that state. Transitions were available for four of the six state-pair combinations (not wake to REM or REM to NREM because they are not commonly observed). Data from all sites were utilized ($n = 45$ implantation sites, 9 animals). The protocol described below was used for both transitions and flickers.

Unlike sustained states (wake, NREM and REM), flickers are very brief and rare. Therefore, to compare them, we needed to develop a sampling procedure that would subsample sustained states as if they were also brief and rare events. A sampling procedure was used to extract a random subset of ISIs from sustained states (for example,

wake for a wake-to-NREM flicker). Two major temporal differences between sustained states and flickers need to be considered. First, there are almost always fewer instances of flickers than instances of sustained states. Explicitly, each individual instance of a flicker is generally bookended by two instances of a sustained state, one before the flicker and one after (that is, a wake-to-NREM flicker is surrounded by wake on either side). Further, each sustained state contains two types of flicker within it (for example, NREM-to-wake and NREM-to-REM flickers both occur within sustained NREM). Second, instances of sustained states are almost always orders of magnitude longer than instances of a flicker (that is, there are many more ISIs during an instance of sustained wake than during a wake-to-NREM flicker).

Our method accounts for these differences by extracting a subset of ISIs from surrounding states that are the aggregation of one short sequence of contiguous ISIs for each flicker, where the number of ISIs in the sequence is equal to the number of ISIs observed during the corresponding flicker. So, for example, if we found two NREM-to-REM flickers (one containing three ISIs and one containing two ISIs), we would similarly sample three consecutive ISIs and separately two consecutive ISIs from random time points during sustained NREM as a control subset. ISIs that overlap with the edge of the flicker are considered to occur during the flicker, so long as the ISI is not far longer than the length of the flicker (>3 s, where an average flicker is ~200 ms).

We then randomly resampled ten ISIs from this subset for each state and flicker type. In the case of flickers, this aggregates ISIs across multiple flickers, because on average a flicker will only have one to two ISIs, which is an insufficient sample for state estimation. We repeat this 100 times for each flicker type. For each sample of a neuron's spiking (ten ISIs), we calculated the mean ISI of the neurons.

Firing rate directional perturbation (Fig. 6d)

To evaluate the directional perturbation of portions of the population, we used the same sampled data used in PCA. For each flicker type, we iterated through each of the 100 samples for the sustained state, flicker state, transition state and predicted state in parallel. For example, for the wake-to-NREM flicker state, wake is the surrounding state, natural transitions from wake-to-NREM (falling asleep) is the transition state, and wake is the predicted state. Iteratively, for each neuron, for each state, we identified whether the neuron's instantaneous firing rate (inverse of mean ISI) was elevated relative to the sustained state. For each sample, we evaluated the percent of neurons whose firing rate was increased (by any amount) relative to the sample of the surrounding state. As a negative control, we compared intervals of the sustained state to each other (sample 1 to sample 100, sample 2 to sample 99, and so on). For this negative control, the portion of units with increased firing should ideally be ~50% (representing chance); however, in the case of some rare flicker types (where few ISIs could be sampled for the controls), it deviated from this due to sampling error.

PCA (Fig. 6f). Data from all sites were utilized, except one site that was almost entirely multiunit (yielding only one single unit, which is insufficient for PCA): $n = 44$ implantation sites, 9 animals. The data were organized into a matrix with a number of columns equal to the number of spike-sorted neurons and rows equal to the number of spiking samples across states. Generally, 100 × 4 samples were available (number of samples for sustained state, flicker state and predicted state). However, if transitions were not available for that flicker type (for example, wake-to-REM flickers are observed but wake-to-REM transitions are not), 100 × 3 were used. The neuron dimension of this matrix is reduced by PCA. Before PCA, for each flicker type, we z-score normalized the sampled mean ISI for each neuron in the performed PCA. This z-score normalization was fit to the samples from the surrounding state and the predicted state (for example, for a wake-to-NREM flicker, wake and NREM appropriately). The inverse of this—mean instantaneous firing rate—gave similar results for this analysis.

PCA was fit to the samples of the sustained state and predicted state to capture the major sources of variance between these two states. PC1 consistently separated these two classes with little overlap. We then transformed samples from the appropriate flicker type and transition onto this axis to observe how they exhibited these sources of variance. Transitions were available for four of the six state-pair combinations (not wake-to-REM or REM-to-NREM because they are not commonly observed).

To enable us to evaluate trends across clustering blocks, the surrounding state and predicted state varies as a function of many factors which cannot be held constant across recorded regions/animals (for example, the number of spike-sorted neurons). Therefore, we MinMax scaled the projections of samples along PC1 to make the positions relative and more comparable across recordings.

We make a technical point about interpreting this rescaled axis here. On this rescaled axis, the distribution of transitions and flickers take an intermediate value near 0. In the original PCA spaces of individual recordings, samples of random noise would project at the origin. However, the 0 point on the rescaled axis corresponds to the middle of the range of projected samples for a recording, and not the origin. Therefore, the intermediate points of flickers and transitions (near 0) along the rescaled axis are not consistent with the projections expected from noise. For confirmation, for each recording, we generated random Gaussian noise from the distributions of firing rates for each neuron in the predicted and surrounding states. We then projected it into these PCA spaces in the same way flickers and transitions are projected into it. For each flicker type, in 70–80% of recordings, the mean projection of the samples is at the origin (with a margin of the standard error of the mean (s.e.m.)). This is not 100% simply because of sampling error; we have limited observations of rare events (flickers) and account for that in our methods. In contrast, the mean projection of the flicker samples is very rarely at the origin (<10% of recordings). Similarly, sampling error probably contributes to this value being nonzero. Therefore, the projections of flickers and transitions are highly different from projections of random noise ($P < 0.001$, t-test) and should not be interpreted as random noise.

PCA projections were multiplied by −1 if the mean value of the projections of the predicted state was less than the mean value of the projections of the surrounding state. This aided visualization by ensuring samples of the surrounding state took on negative values and samples of the predicted state took on positive values.

Error bars were challenging to formulate due to ISI resampling and possible dependency between spiking in two regions during co-flickering. We sought not to incorporate resampling into this calculation (because it would lead to an artificial overestimate of our precision). Therefore, we took the mean projection of all samples for each surrounding state, predicted state, flicker state and transition state in a clustering block. The error bars are computed from only these values (the number of data points is equal to the number of recordings). We implemented an error metric that is intuitively similar to standard deviation, with a focus on capturing animal-wise variance. The s.e.m. was multiplied by the square root of the number of animals (that is, the number of uncorrelated observations). This gave error bars that were roughly reflective of the trends seen in independent statistical significance tests.

For each flicker type, we fit a linear mixed effects model relating state to PC1 projection with animal as a random effect. We then performed an analysis of variance (ANOVA) and post-hoc EMMeans with Tukey correction. When grouping flickers by flicker type, we multiplied P values by 6 for Bonferroni multiple hypothesis correction. When grouping flickers by region, we multiplied P values by 10.

Flickers during waking behavior (Fig. 7b,e,h and Extended Data Fig. 4g). To correlate animal behavior with flickers, we aligned video (15 fps) with neural data using a dedicated digital data channel

synchronized with electrophysiological data acquisition, providing nanosecond-accurate timestamps per video frame (E3Vision, WhiteMatter, LLC). We used Farnebäck dense optical flow[80] to measure movement. Specifically, motion was measured using dense optical flow that provides a per-pixel movement vector calculated between temporally adjacent video frames. Dictated by the video sampling rate, 15 per-pixel optical flow vectors were produced per second of recording. A video frame is segmented into three notable parts: (1) the subject animal, which is the core focus, (2) the background that should be excluded and (3) the headstage tether cable that should be excluded but produces the highest movement vectors due to its contrast with the background, immediate proximity to the camera and quick jittery motion. To reduce the effect of cable movement, we excluded the top 3% of motion vectors. To focus analysis on the animal (as opposed to the background), we took the remaining top 10% of motion vectors. Explicitly, this retained the 87th percentile to 97th percentile of motion vectors produced by dense optical flow. Total animal movement was computed as the mean of these motion vectors for each frame.

We next calculated two normalizations to allow comparisons across animals and over time. To accommodate drift and environmental changes, such as light/dark, we rescaled all motion vectors per each hour of video to a [0, 1] range. To normalize differences between mice, we computed the 75th percentile of motion vectors per mouse and rescaled the movement values associated with each frame such that each animal's 75th percentile movement values were aligned. Finally, we collapsed movement values above 1 to 1, resulting in a movement vector in the [0, 1] range. In other words, periods of high movement appeared as a sustained sequence of 1 s. These normalization steps produced a movement value that is sufficiently invariant to light/dark cycles, differences in recording resolution, and variations in camera orientation and zoom to align locomotor states across animals.

Three substates of waking were defined: sustained high activity, low to intermediate activity, and brief pauses embedded within high activity. We computed two median filters over the normalized per-video frame movement values. Median filters are particularly useful in this context because they produce a smoothing with a sharp contour to the data that is cleanly delineable. The first was a rolling median using a 60 s window. At a threshold of 0.75, this median filter broadly segmented periods of sustained, high activity and periods of intermediate to low activity. Within periods of high activity, we then used a rolling median using a 0.66 s window to segment pauses from within sustained high activity. We manually evaluated a random subset of each locomotor state to confirm the accuracy of the algorithm parameters.

When plotting the time course of flicker rate in relation to a motion state, we plot the rate of flickering for two motion states. One of these motion states is observed to be brief (for example, pause) and occur amid a generally longer state (for example, active) that is observed to frequently surround the shorter state. The time course of any motion state is variable; therefore, we scale the length of the state into arbitrary units between values of 0 (start of the state) and 1 (end of the state). We divide this time course into 20 bins (that is, bin 1 is 0–0.05,..., bin 20 is 0.95–1) and calculate the mean rate of flickers during each bin. We plot the second half of the longer state first along the *x* axis (for example, bins 11–20 of active), then we plot all bins of the shorter state (for example, bins 1–20 of pause), followed by the first half of the longer state (for example, bins 1–10 of active). We do not smooth these values; the curving in the lines arises from using the scipy.interpolatemake_interp_spline function.

Statistical comparisons of flicker rate use one value of the mean flicker rate across the span of a continuous bout of a state (that is, the number of flickers which occurred during a single pause normalized to the duration of the pause). Statistical comparisons of flicker duration take one value per flicker (the length of that flicker).

Flickers during sleeping behavior (Fig. 7a,c,d,f,g,i). Using an analogous process to wake, three substates of sleep were defined: sustained high activity (indicative of an extended arousal), low activity and brief high activity embedded within low activity (indicative of a twitch).

To capture minor fluctuations in optical flow (indicative of brief stillness, or 'freezing'), we inverted the trace of the magnitude of the optical flow and applied the scipy.find peaks function with default hyperparameters. We identified freezing as the 1 s following each of these peaks. In contrast to other analyses where the bookending state is an extended motion state, here the bookending state is NREM. This is suitable because these freezing events are specifically examined within the NREM state.

Evaluating contribution of common artifacts (Extended Data Fig. 9). Given the strong correlation of the CNN's classification with behavior, we sought to confirm that its predictions did not arise from common noise artifacts (particularly muscle activity). We removed these artifacts by subtracting the common mean of all probe channels from the broadband signal (Extended Data Fig. 9a). We then tested our pretrained CNNs on this denoised input. Because of the precise pattern recognition employed in deep learning, any ablation of the input after training has the potential to put the input out-of-distribution for the CNN. Still, we observed that in the majority of models the accuracy was uncompromised (Extended Data Fig. 9b)—a strong indication that models learn independently of such noise. We also sought to confirm that the general pattern of errors (such as flickers) was not affected by this denoising. We observed that the pattern of errors in general, and specifically around flickers, is preserved (Extended Data Fig. 9c–e). In Extended Data Fig. 9d, a cross-correlogram was calculated by comparing the time series of predicted probability for the flicker state (that is, NREM for a wake-to-NREM flicker) in a 60-s window surrounding each flicker. For each probe, these flicker-wise time series were concatenated. Afterwards, for each lag, the signal of interest was cyclically shifted, and the Pearson correlation coefficient was calculated between the original probabilities and the shifted mean-subtracted signals. This process was first applied both to the actual data to obtain the experimental correlations and second to randomly permuted versions of the same data to generate control correlations. The actual data produced strong correlations that were easily distinguishable from random control fluctuations.

Software and statistical analyses

Data are reported as mean ± s.e.m. unless otherwise noted. We utilized mixed effects regression analysis, treating animal as a random effect, followed by EMMeans with Tukey or Bonferroni post hoc correction (lme4, R)[81] to determine statistical significance ($P < 0.05$). Where appropriate, the normality of residuals was assessed using the Shapiro–Wilk test and visually confirmed through quantile–quantile plots. Plots of residuals versus fitted values were examined to ensure homogeneity of variance across predicted values. Model selection was guided by comparing Akaike's information criterion and Bayesian information criterion, with ANOVA from base R used to compare models and assess the significance of fixed effects, taking into account the random effects structure. The significance of fixed effects was evaluated using Wald chi-square tests. All statistical tests were two-tailed with an alpha level set at 0.05.

When appropriate, a log transformation was applied to data for normalization. In some cases (for example, Supplementary Fig. 4), the normality of data distributions was not tested.

Note that, in some cases, CNN models that failed (that is, dropped to 33%, picking only a single class) were excluded from analyses, as these imposed a bimodality on the data. Such instances are noted in the main text.

All software used in this study was written in Python and R, and all auxiliary code was sourced from open repositories.

Reporting summary
Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability
The datasets generated and/or analyzed in this study constitute tens of terabytes of raw neural broadband. The data are stored in a cost-efficient manner not immediately accessible to the internet. We are excited to share data upon reasonable request and as technical limitations make possible. The Allen Brain Atlases used for stereotaxic targeting are available at https://atlas.brain-map.org/.

Code availability
All relevant code from our lab, including software needed to run recordings or CNN models like ours, is in Python 3 and is publicly available via GitHub at https://github.com/hengenlab. Other groups' code, including Open Ephys, SpikeInterface and MountainSort4, is publicly available as specified in Methods.

References

72. Siegle, J. H. et al. Open Ephys: an open-source, plugin-based platform for multichannel electrophysiology. *J. Neural Eng.* **14**, 045003 (2017).
73. Buccino, A. P. et al. SpikeInterface, a unified framework for spike sorting. *eLife* **9**, e61834 (2020).
74. Science: Public Resources: Atlases: Allen Mouse Brain Atlas. *Allen Institute for Brain Science* http://www.alleninstitute.org/science/public_resources/atlases/mouse_atlas.html (2012).
75. Mathis, A. et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
76. Good, I. J. Rational decisions. *J. R. Stat. Soc. Ser. B* **14**, 107–114 (1952).
77. Brodersen, K. H. et al. The balanced accuracy and its posterior distribution. In *20th International Conference on Pattern Recognition* 3121–3124 (IEEE, 2010).
78. Kelleher, J. D., Namee, B. M. & D'Arcy, A. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies* (The MIT Press, 2015).
79. Siapas, A. G. & Wilson, M. A. Coordinated interactions between hippocampal ripples and cortical spindles during slow-wave sleep. *Neuron* **21**, 1123–1128 (1998).
80. Farnebäck, G. *Two-Frame Motion Estimation Based on Polynomial Expansion. Lecture Notes in Computer Science* (ed. Goos, G.) 363–370 (Springer, 2003).
81. Bates, D. et al. Parsimonious mixed models. Preprint at https://arxiv.org/abs/1506.04967 (2015).

Acknowledgements
This work is supported by NIH BRAIN Initiative 1R01NS118442-01 (KBH), and the Schmidt Futures Foundation SF 857 (D.H.). Through the Pacific Research Platform, this work was supported in part by NSF awards CNS-1730158, ACI-1540112, ACI-1541349, OAC-1826967, the University of California Office of the President, and the University of California San Diego's California Institute for Telecommunications and Information Technology/Qualcomm Institute. We acknowledge CENIC for the 100 Gpbs networks. We also thank S. Aton, B. Carlson, S. Ching, C. Cirelli, M. Frank, K. Ganguly, T. Holy, L. de Lecea, D. Redish and P. Shaw for their insights and conversations about this research.

Author contributions
D.F.P. developed and ran the models and performed behavioral analyses, A.M.S. performed statistical analyses, wrote the paper, performed the single-unit spiking analyses and performed behavioral analyses, Y.X. provided sleep-scoring expertise and animal care, S.J.B. contributed substate analyses, S.F. provided sleep-scoring expertise, D.T. performed flicker identification, T.B. provided intellectual and technical consultation, E.L.D. provided mentorship and consultation, D.H. provided mentorship and consultation and K.B.H. led, directed and envisioned the project, edited figures and wrote the paper.

Competing interests
The authors declare they have no competing interests.

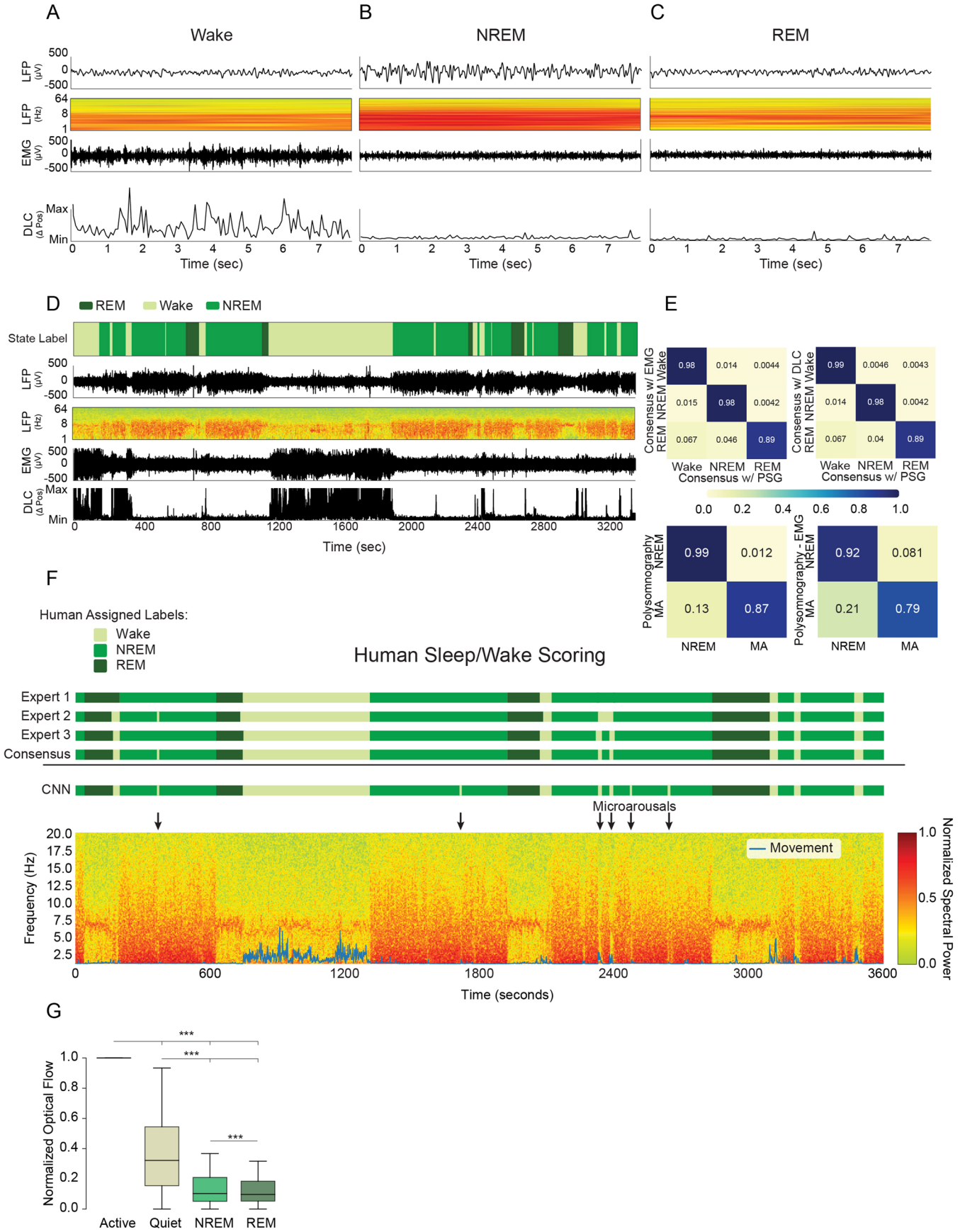
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41593-024-01715-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-024-01715-2.

Correspondence and requests for materials should be addressed to Keith B. Hengen.

Peer review information *Nature Neuroscience* thanks Tatiana Engel and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41593-024-01715-2

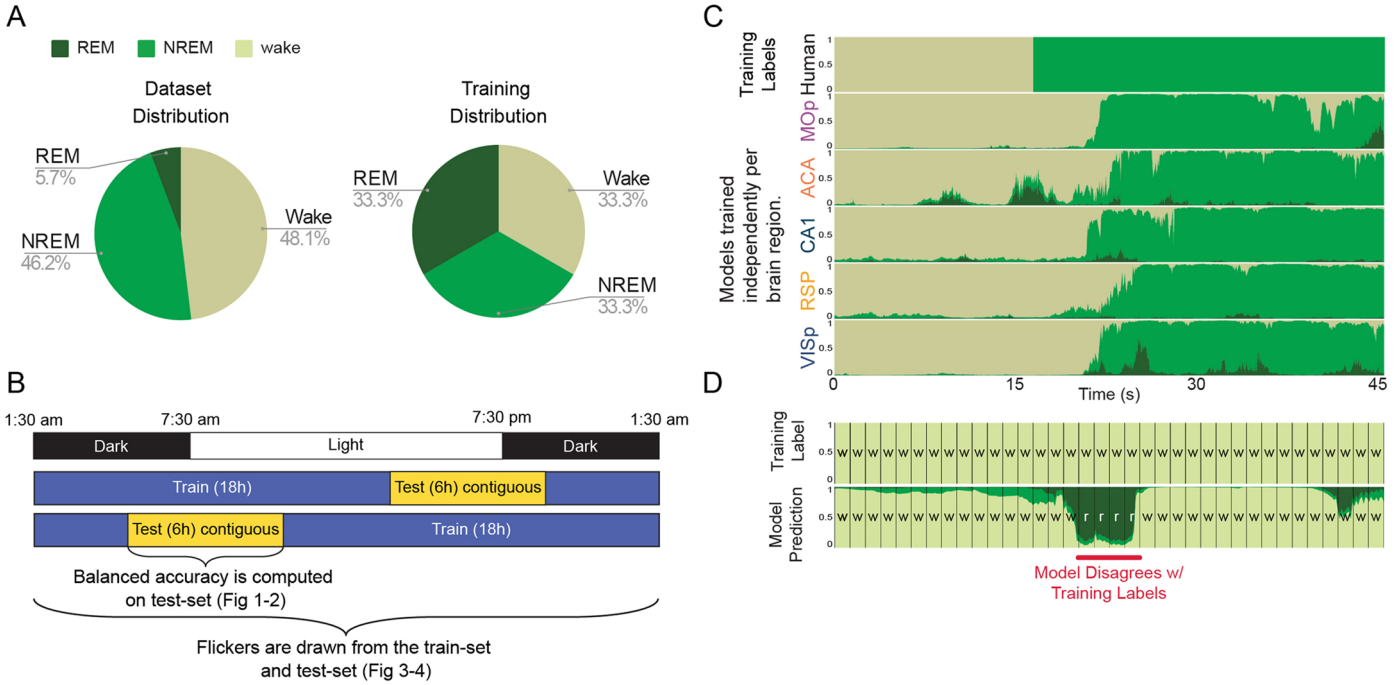


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41593-024-01715-2

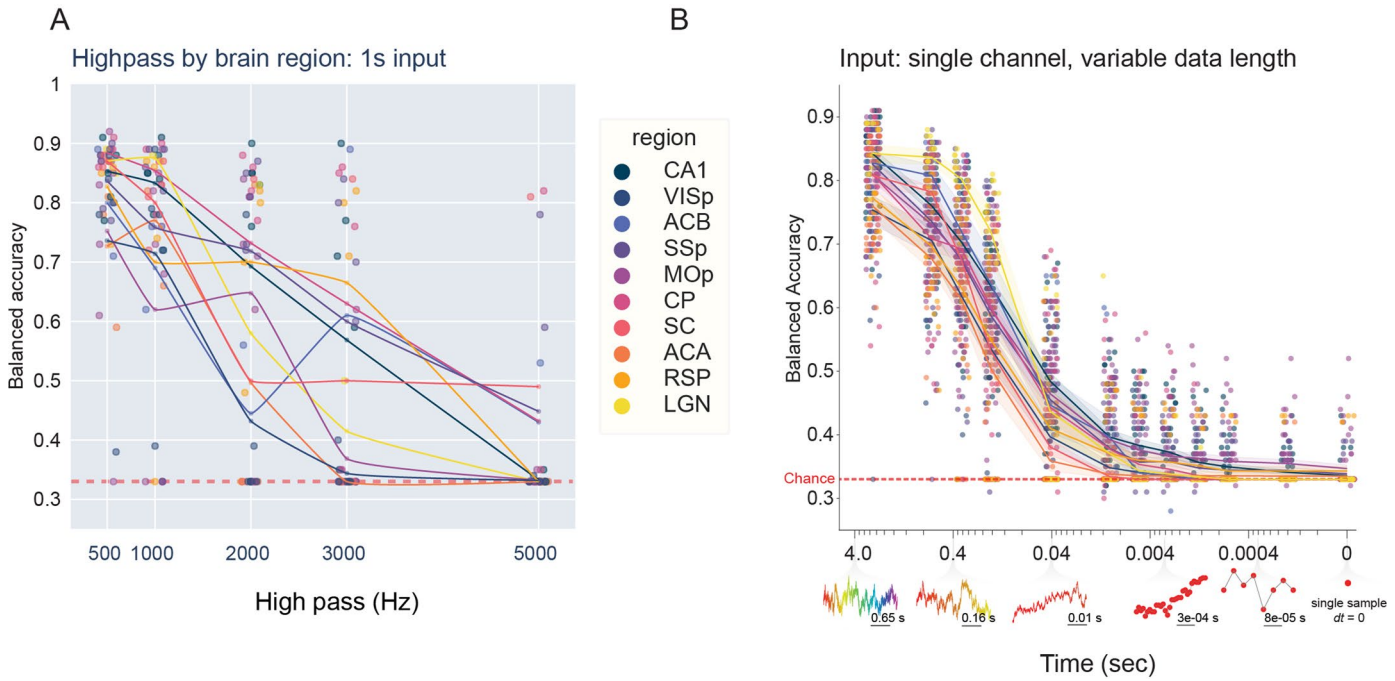
Extended Data Fig. 1 | Human sleep scoring, polysomnography, example traces. A, 8-seconds of exemplary polysomnography data from an animal during wake. From top-to-bottom: the low-passed LFP trace, the LFP spectrogram, the EMG, and the MinMax normalized change in position tracked by DeepLabCut (DLC) are shown (Mathis et al.[75]). To form EEG-like traces, raw data were low-pass filtered at 125 Hz, downsampled to 500 Hz, and 8 non-adjacent channels were averaged. **B**, 8-seconds of exemplary data from NREM, **C**, 8 seconds of exemplary data from REM. **D**, An hour of data sleep-scored for Wake, NREM and REM with polysomnography. Delta band power (0.1-4 Hz) is highly enriched in NREM (slow wave) sleep, theta band power (6-8 Hz) is enriched in REM sleep, and high resolution motor output disambiguates periods of waking, including microarousals (Watson et al.[57]). **E**, *Top*- Confusion matrices for average individual scorer performance using EMG or DLC for motion (rows) to consensus scoring among three expert scorers using polysomnography, including DLC and EMG (columns). *Bottom*- Confusion matrices for microarousal detection in NREM for average individual scorer performance using polysomnography (including DLC and EMG) and polysomnography (with DLC, but not EMG). Ground truth is consensus of three expert scorers using polysomnography, including DLC and EMG (columns). **F**, Exemplary hour of data sleep-scored by three experts to produce a consensus score, which is used as training data for a CNN (predictions shown). Blue movement trace is based on DeepLabCut. **G**, Box plots of normalized optical flow (see methods section "Flickers During Waking Behavior") for active wake (active motion state), quiet wake (inactive motion state), NREM, and REM. Box plots box the intermediate quartiles, and whiskers extend 1.5 IQR beyond this box. Highly significant differences in means were observed for all comparisons ($p < 0.0001$ for all comparisons, ANOVA with post-hoc EMMeans with Tukey correction, n = 1 animal). *ns* $p > 0.05$, * $p <= 0.05$, ** $p < 0.01$, *** $p < 0.001$.

https://doi.org/10.1038/s41593-024-01715-2



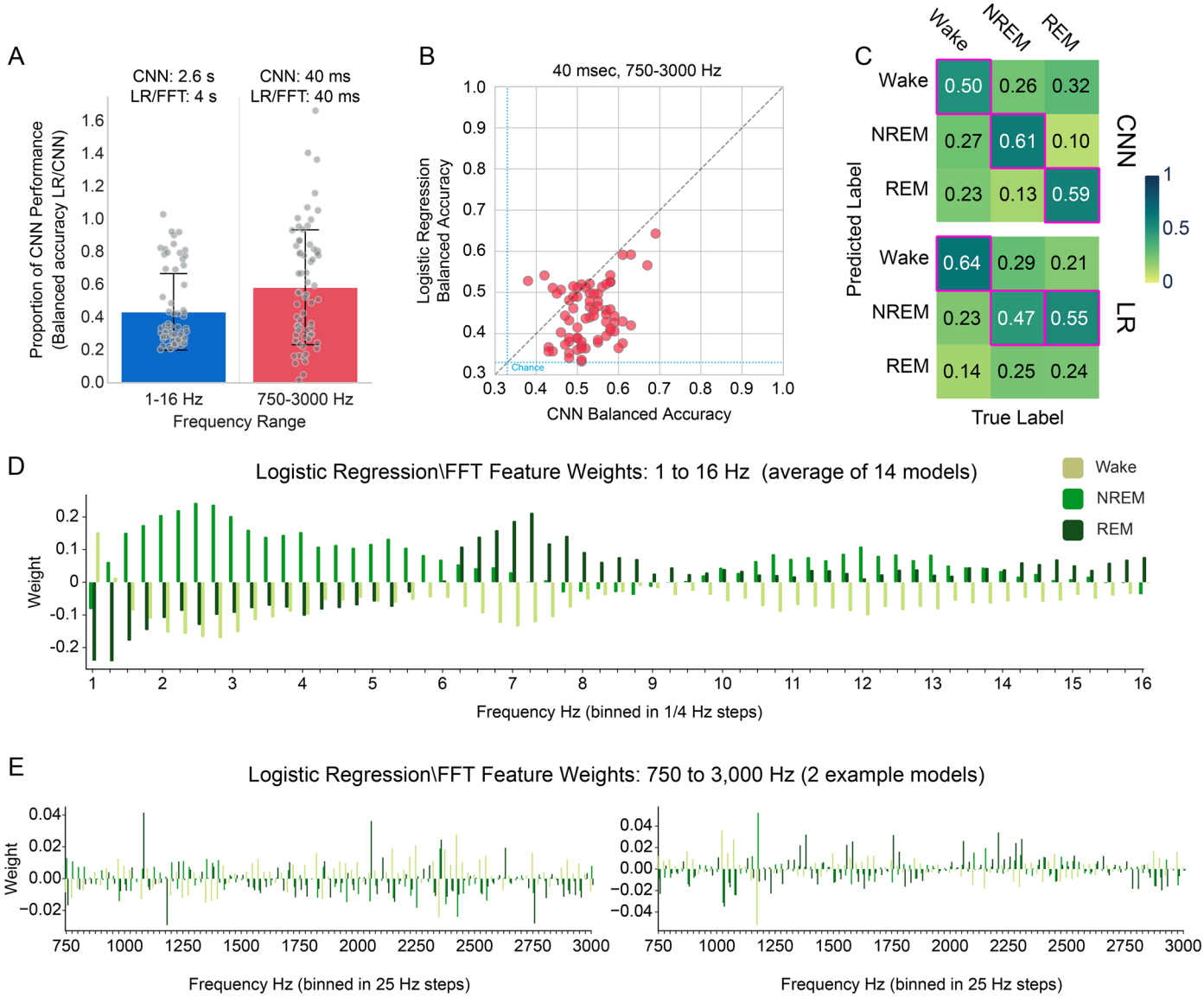
Extended Data Fig. 2 | CNN train/test separation, model independence, and flickering. A, The three arousal states, REM, NREM, and wake, were not evenly distributed in the recorded data (dataset distribution). Training on equal amounts of data from each state, that is, class balancing (training distribution), prevents CNN models from simply learning to predict the most frequent class whenstate information is not clear. **B**, 24 h of data spanning a complete 12-12 light/dark cycles was included from each animal. Cartoon is a schematic illustrating 24 h of light/dark (top row), and two examples of train/test segmentation of data (bottom rows). CNNs were trained on 18 h of data. Test data comprised six contiguous hours of data spanning a light/dark transition. **C**, Models were trained using consensus human labels (top row). Independent models were trained and tested in each brain region within each animal (bottom five rows show the predictions of five models, each trained to recapitulate the human labels from data recorded within the brain region indicated on the left). CNNs are well suited to overcome label error, such as when a human score lags or leads a transition. The provided example demonstrates disagreement and model independence surrounding a global state transition (wake into NREM). The y-axis of CNN models indicates instantaneous confidence [0-1] in each state via the proportion of each of three colors at each point in time. **D**, In contrast to assessment of model accuracy, flickers are extracted from both train and test output. Flickers detected in the training component of data represent CNNs directly disagreeing with training labels (red line indicates example of wake-to-REM flicker). Human labeled data (top row) and corresponding CNN predictions (bottom row) are computed on an interval set by the experimenter. Flickers were extracted and cross validated in both 2.6 s and 1 s interval models.

https://doi.org/10.1038/s41593-024-01715-2



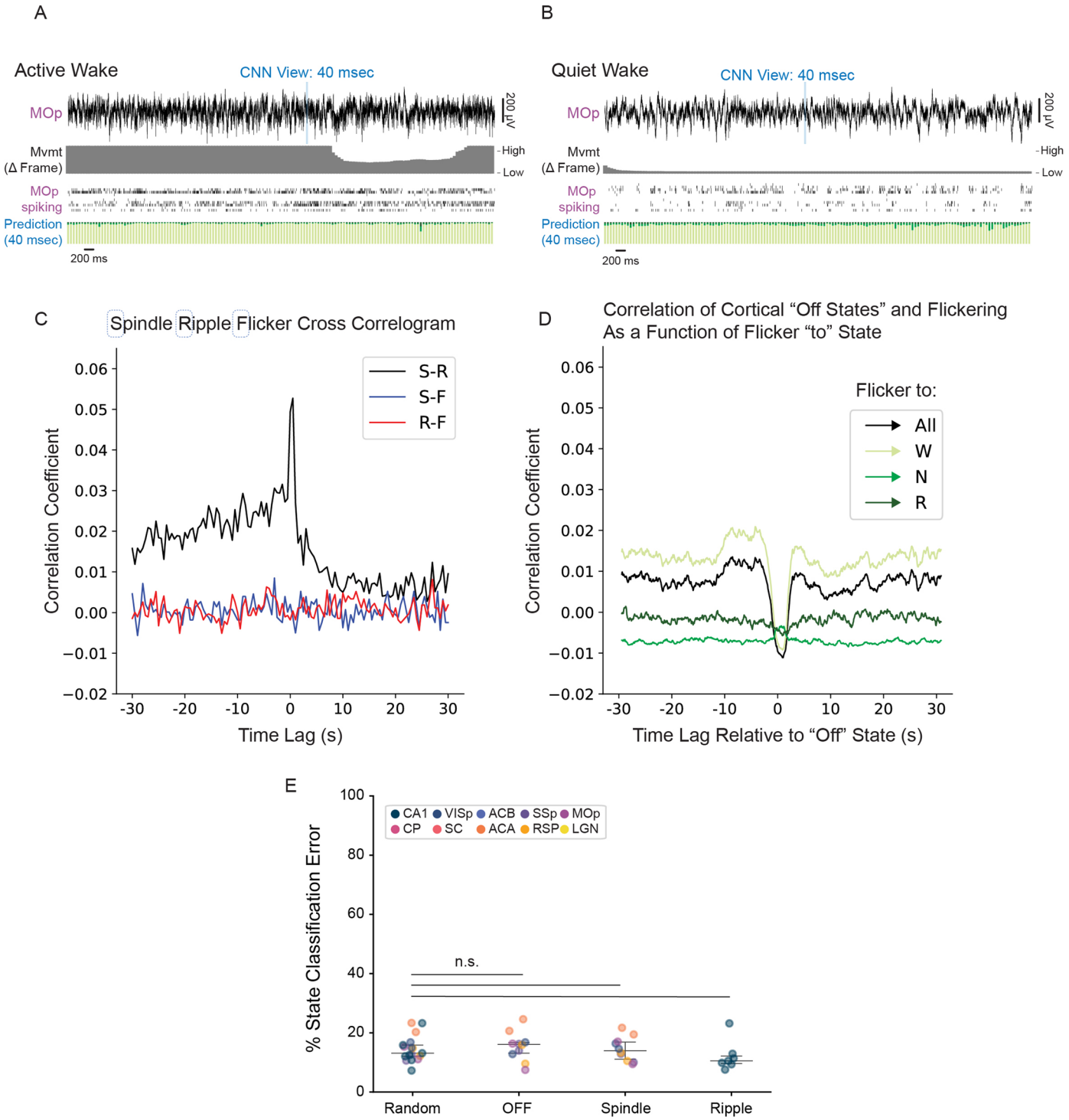
Extended Data Fig. 3 | CNN accuracy by region as a function of high-pass and sample size reduction. A, Balanced accuracy of CNNs trained and tested on progressively high-passed raw data from each recorded brain region (n = 45 implants, 9 animals, 10 regions). Dot color represents the region from which this model was trained. The region-colored line traces the average balanced accuracy of models trained on data from that region across various levels of high-pass. High-pass filtering significantly decreased brain state information above 1,000 Hz (y = -9.892e-05*x + 0.84, p < 0.0001, r2 = 0.43). **B**, To directly test the minimum time interval in which sleep and wake states reliably structure neural dynamics, we trained and tested a series of CNNs on single channel data, each model operating on a progressively smaller interval of data (from 2.6 s to 0 s). Accuracy declined as a function of number of input sample points (pearson correlation: r = 0.650464, p < 0.0001). Example data at various input sizes is shown below the x-axis. Model accuracy is shown as a function of region (marker color). Regional means are shown in colored lines (+/- SEM, shaded area).

https://doi.org/10.1038/s41593-024-01715-2



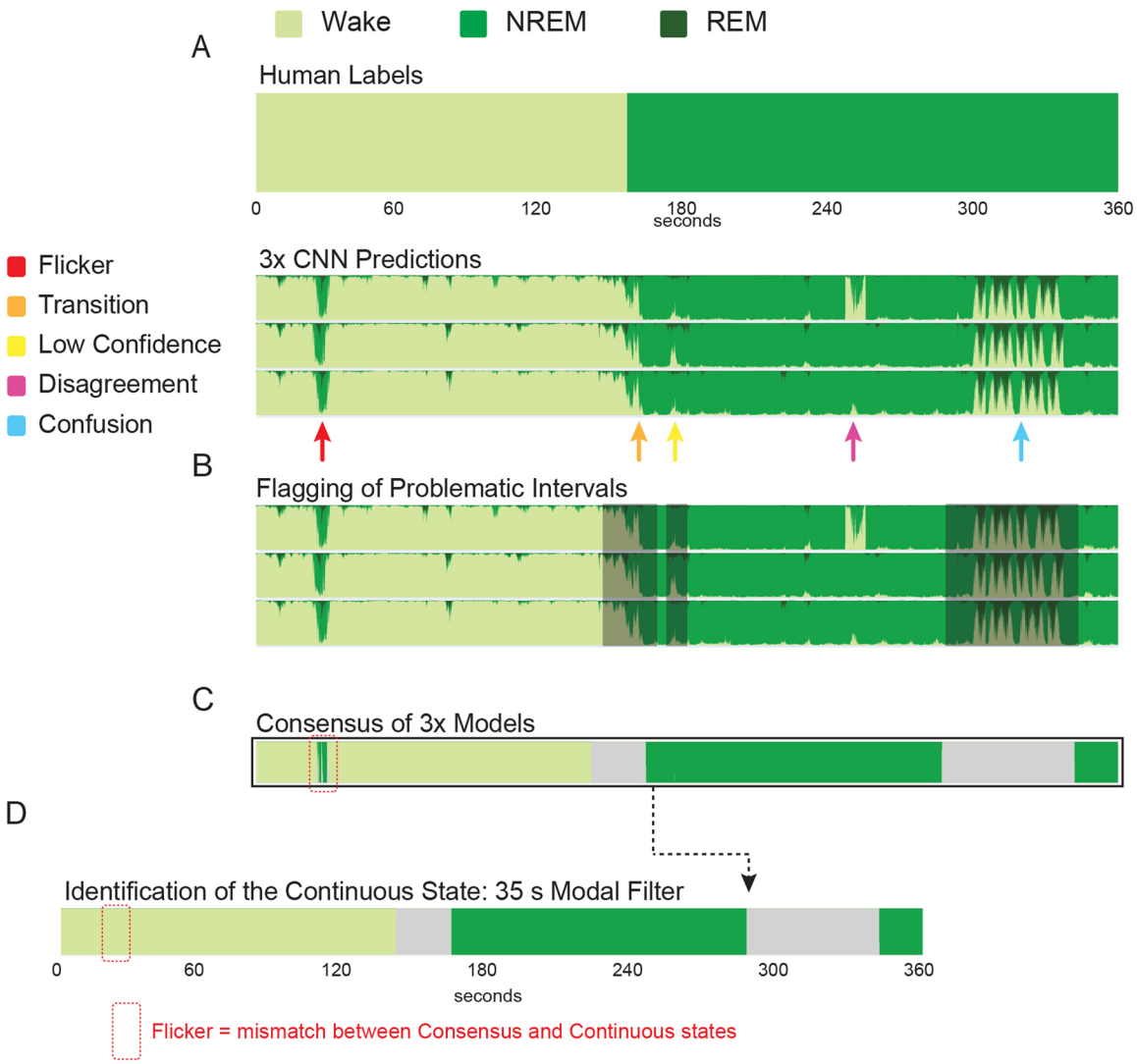
Extended Data Fig. 4 | Spectral density does not capture the full extent of high frequency state embedding. To evaluate whether the embedding of brain state in > 750 hz activity is explained by spectral bandpower, a logistic regression (LR) was trained on either low frequency bandpower (1-16 Hz) or high frequency bandpower (750-3,000 Hz). The resolvability of state in the LR was compared to the performance of a CNN exposed to the same frequencies. Analyses consider the 71 single channel models that contribute to data in Fig. 3. **A**, Proportion of the CNN's balanced accuracy that is achieved by LR. Data are presented as mean values +/- SEM. At both low frequencies (1-16 Hz blue), and high frequencies (750 - 3,000 Hz, red) the CNN substantially out performs a LR model using bandpower/FFT features. The difference between LR vs. CNN models is more pronounced in the 750 - 3,000 Hz range with the LR performance slightly better than ½ the balanced accuracy of CNN models. **B**, Unity line plot comparing the performance of 750 - 3,000 Hz LR versus 40 ms CNN models on a channel-by-channel basis. Note that points generally fall below the unity line, indicating higher balanced accuracies with the CNN than the LR. **C**, Confusion matrices for a single channel's models - top is the CNN, bottom is the LR. Note that the CNN learns information about all three states - the diagonal is characterized by above chance accuracy (white lettering, red boxes). In contrast, the LR collapses to a two-state solution (sleep v wake). For A-C, n = 69 individual channels, where each implant is represented at least once. **D,E** To understand the frequency-based patterns that contain reliable state information, we examined the feature weights for the low and high frequency LR models. **D**, 14 high-accuracy 1 - 16 Hz LR model weights were averaged together. The signatures of state learned in a data-driven fashion are, unsurprisingly, highly consistent with human heuristics. For example, the 1-4 Hz (delta) band is highly weighted during NREM, and the 6-8 Hz (theta) band is highly weighted during REM. **E**, High frequency LR models did not exhibit consistent patterns across channels. Here we show two exemplary 750 - 3,000 Hz LR models. While each model learned to score state at > 50% balanced accuracy, the two weight distributions show distinct learning patterns, characterized by various irregularly interspersed frequency bands. This supplemental analysis employing logistic regression models trained on bandpower features provides valuable insight into the spectral components of dynamical signatures of sleep and wake states. While a CNN-based approach confirms the critical role of high frequency patterns in state classification, logistic regression analyses offer a complementary perspective, suggesting that state information persists in the high frequency domain, albeit less comprehensively than captured by CNNs. This dual approach underscores the complexity of neural embedding of states and the limitations of relying solely on traditional frequency domain analyses to understand such dynamics.

https://doi.org/10.1038/s41593-024-01715-2



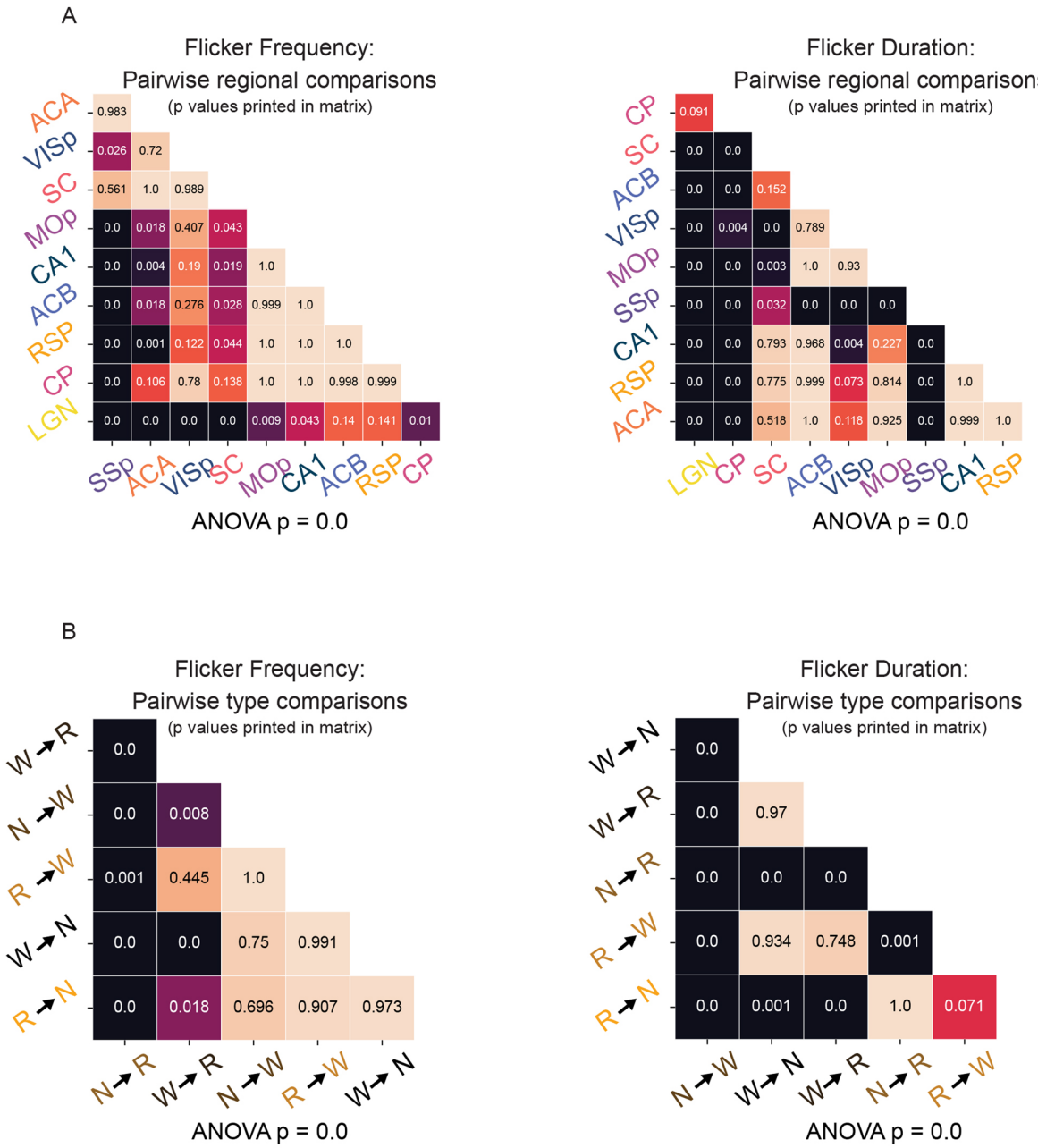
Extended Data Fig. 5 | 40 ms CNNs are resilient to oscillatory substates. **A**, Top- Broadband trace of exemplary activity during active wake in MOp over several seconds. Blue box shows the width of an individual input data used by the 40 ms CNN to predict state. Middle- Raster of MOp spiking. Bottom- Stacked barplot of CNN prediction probabilities across the three states every 1/15 s. **B**, Exemplary data from MOp during quiet wake. **C**, Cross-correlogram between spindles (S), ripples (R), and flickers (F) in Animal 5. A strong central peak in the cross-correlogram is observed between spindles and ripples consistent with prior work (Siapas & Wilson[79]). No substantial positive correlation is observed with flickers by spindles or ripples. **D)** Cross-correlogram between OFF-states and flickers of various states (for instance ->W includes NREM-to-wake, and REM-to-wake). A major central correlation trough is observed in flickers to wake, meaning flickering is reduced during sleep OFF states. **E)** Percent of substates which coincide with errors in model classification in 1 s CNNs. Stacked black lines show the intermediate quartiles with all points as swarm scatter colored by region. Cortical OFF states and sleep spindles were detected in all cortical regions in two animals (n = 10). Ripples were detected in all recordings of CA1 hippocampus (n = 7). As a negative control, for each substate, an equal number of randomly selected timestamps were selected and evaluated. No significant differences ($p > 0.05$) were found between this negative control and any of the three substates.

https://doi.org/10.1038/s41593-024-01715-2



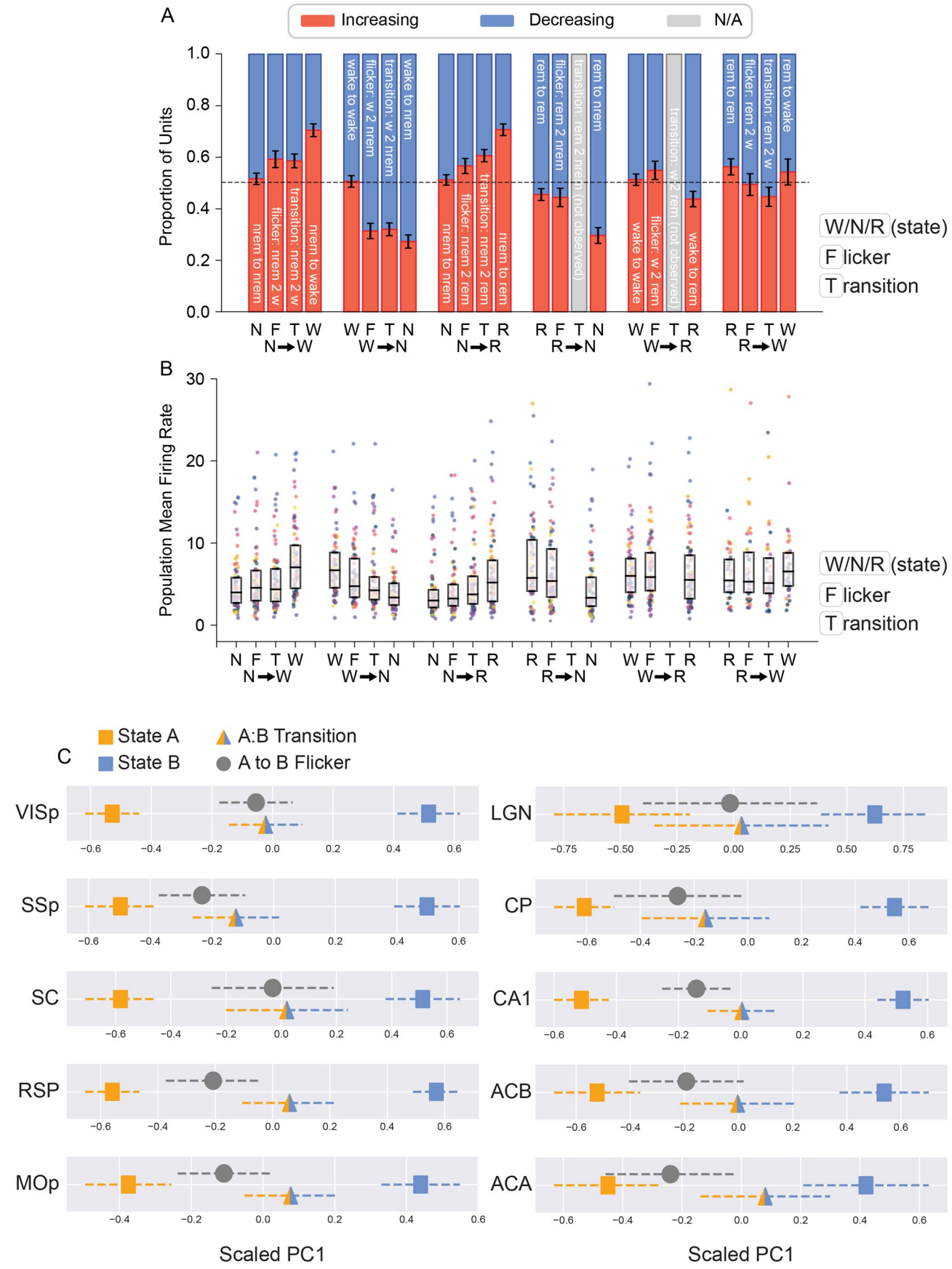
Extended Data Fig. 6 | Flicker definition method applied to synthetic data. To illustrate the process of identifying flickers within data, we generated a synthetic dataset equivalent to 64 channels of recording from a single brain region. The synthetic data were designed to contain examples of all forms of noise that our algorithms exclude in the process of identifying flickers. **A**, Human experts score animal arousal state based on polysomnography (top). Based on these scores, three CNNs are trained to identify NREM, REM and wake based on raw neural data (1 s input) from all 64 channels within a brain region. Shown are triplicate CNN-generated state scores for our simulated brain region. **B**, We identify and exclude two forms of extended intervals during which flickers are not to be considered. First, we identify windows of time during which a *transition* (orange arrow) between two sustained states occurs. Second, we identify rare instances of label *confusion* (typically the result of extreme noise or CNN error) (blue arrow). This is achieved by automatically excluding any 35 s epochs with a mean confidence <75%. To exclude low confidence *noise* (yellow arrow) from our analyses, we eliminate 1 s epochs with a mean confidence <75%. For each time point, we then assign the label of the most confidently-predicted state. **C**, To exclude *disagreements* (magenta arrow), or artifacts of particular CNN models (that is predicted by only one of three CNNs), we collapse predictions across models by selecting the most commonly predicted state at each time point. **D**, We then slide a 35 s modal filter across the majority state array to create a label corresponding to the stable macro state surrounding each point in time. *Flickers* are defined as disagreements between the consensus array and the continuous array (red arrow).

https://doi.org/10.1038/s41593-024-01715-2



Extended Data Fig. 7 | Flicker timing shows significant differences by circuit and flicker type. Pairwise comparisons between subsets of flickers using EMMeans with Tukey correction as a post hoc-test for an ANOVA fit to a linear mixed effects model with animal as a random effect. For all ANOVAs, $p < 0.001$. p-Values for the pairwise comparison of row and column are printed inside a cell of the diagonal matrices. Values of 0.0 indicate $p < 0.001$. **A**, Left- Significant differences in flicker frequency (that is rate) when grouped by region (that is circuit). Right- Significant differences in flicker duration when grouped by region. **B**, Left- Significant differences in flicker frequency (that is rate) when grouped by flicker type. Right- Significant differences in flicker duration when grouped by flicker type.

https://doi.org/10.1038/s41593-024-01715-2

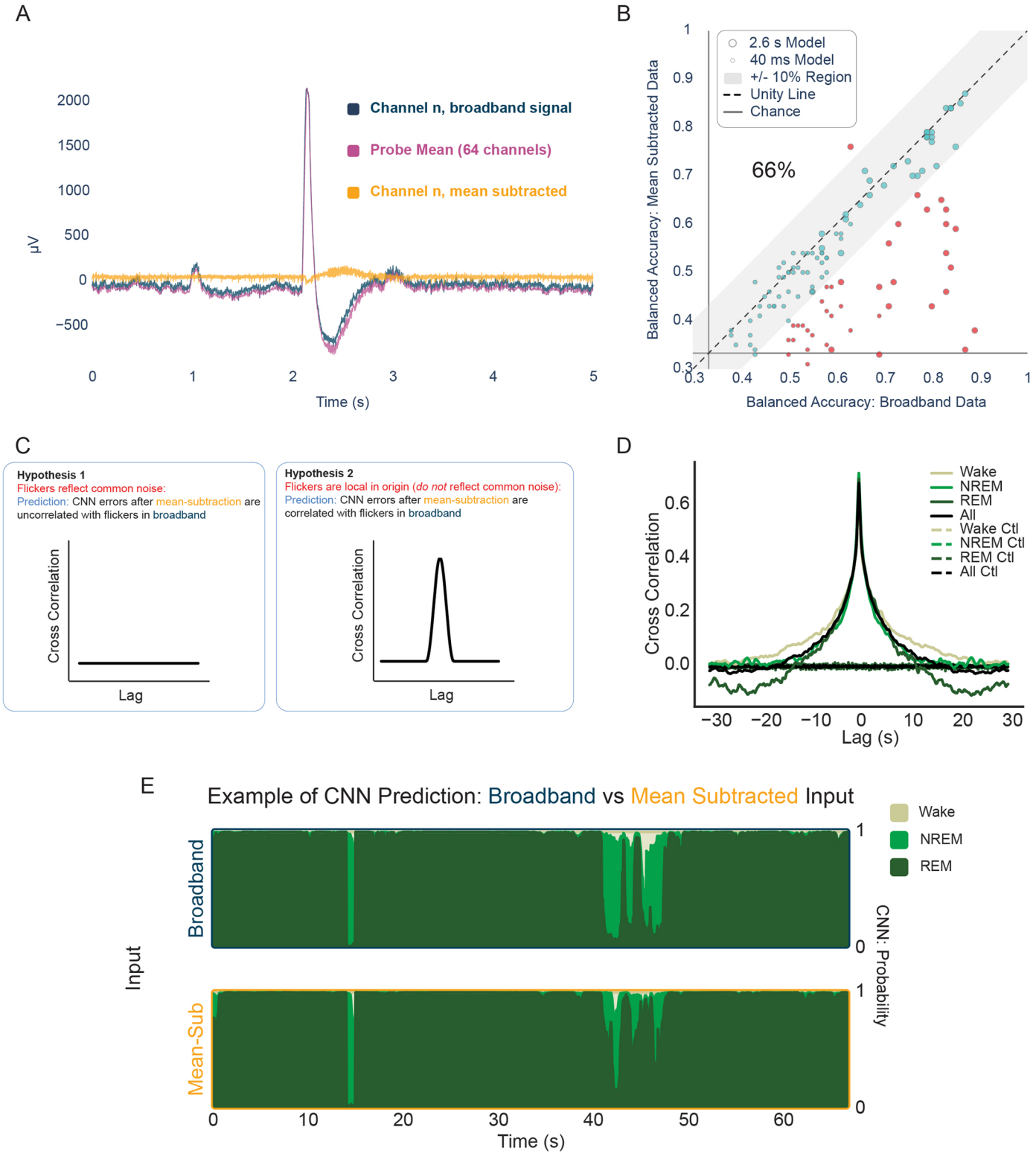


Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s41593-024-01715-2

Extended Data Fig. 8 | Flickering corresponds to transition-like spiking patterns. This figure expands on trends in spiking presented in Fig. 6. **A**, The portion of units whose sampled instantaneous firing rate was different relative to a random sample of the surrounding state: surrounding state vs. surrounding state (negative control), surrounding state-to-predicted state flicker, surrounding state-to-predicted state transition, and surrounding state vs. predicted state (n = 45 implantation sites, 9 animals). **B**, The mean single unit firing rate of recorded circuits during a surrounding state, surrounding state-to-predicted state flicker, surrounding state-to-predicted state transition, and predicted state (n = 45 implantation sites, 9 animals). Box plots show the intermediate quartiles with all points as swarm scatter colored by region. **C)** Mean scaled PC1 projections for the ten regions. For each region surrounding state, predicted flicker state, flicker, and transition are shown. All recorded sites were utilized, except one site which only yielded a one single-unit, which is insufficient for PCA (n = 44 implantation sites, 9 animals). To incorporate the n animals into estimated variance, error bars are the SEM multiplied by the square-root of n animals. See Supplemental Tables 2-5 for significance of pairwise comparisons based on linear mixed effects model projection ~ sample type (that is flicker, transition, surrounding, predicted) + (1 | animal) with post-hoc EMMeans with Tukey correction. Bonferroni multiple hypothesis correction was applied based on the number of regions.

https://doi.org/10.1038/s41593-024-01715-2

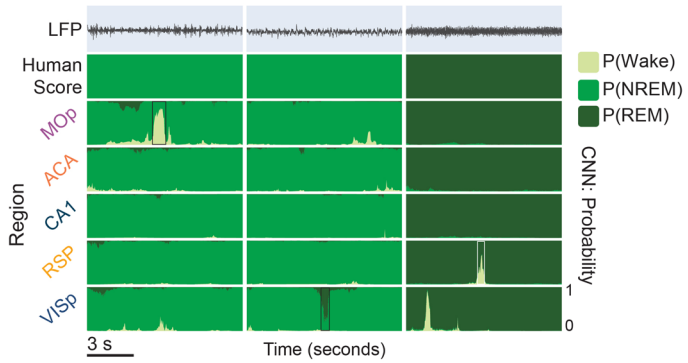


Extended Data Fig. 9 | See next page for caption.

https://doi.org/10.1038/s41593-024-01715-2

Extended Data Fig. 9 | Common signals (eg., EMG, EOG) do not account for the fast embedding of state or flickers. EMG, EOG, and noise events should impact all electrodes within a brain region. To test whether such common signals were the basis of the embedding of state at fast timescales, we evaluated CNN performance and flickers in the context of a mean-subtracted version of our data. **A**, Example raw data (navy) are shown alongside the average signal from the entire array (64 ch; purple). Note that the large, non-neuronal artifact at two seconds is obvious in the array-mean as well as on the single channel. Subtracting the mean from the single channel removes the common artifact (orange), but preserves local spiking (amplitude ~ 50 μV). **B**, We passed the mean-subtracted single channel data into the extant single channel models, which were trained on the broadband raw data. Tested on mean-subtracted data, the majority of models performed within 10% of their balanced accuracy when tested on broadband data (teal points; gray band indicates region of equivalent broadband and mean-subtracted balanced accuracy +/- 10%). This was true for both 2.6 s and 40 msec models (large and small points). Points below the 10% band represent models that either originally learned from shared signals, or models that were disabled by the distortion of the data. Note that the teal points serve as an existence proof that, absent common signals, it is possible to identify brain state in local neural dynamics in 40 msec windows. It is likely that training new models on mean-subtracted data would substantially improve the performance of all models. **C**, To test whether flickers were the result of common signals (such as EMG artifact) we cross correlated disagreements between the CNN and human labels in two contexts: broadband data and mean-subtracted data. Cartoon of the two hypothesis to illustrate the expected results in the null and alternative condition. **D**, Cross correlogram of flickers in broadband and mean subtracted data reveal a peak at zero lag, suggesting that there is a high correspondence of flicker events between the two datasets. **E**, Example CNN predictions on the same time interval in broadband (top) and mean-subtracted data (bottom).

https://doi.org/10.1038/s41593-024-01715-2



Extended Data Fig. 10 | Additional examples of flickering between states. Three exemplary flickers. Flickers are defined as high-confidence, non-global events that are not detected in low-pass models, and are distinct from transitions between states. Top trace is neural broadband. Second row is human scoring of the corresponding state. Bottom five rows are outputs of independent CNNs trained in each of the four brain regions recorded in the same animal. The black box in the left column indicates a NREM-to-wake flicker in primary motor cortex. The black box in the center column demonstrates a NREM-to-REM flicker in primary visual cortex. The white box in the right column demonstrates a NREM-to-wake flicker in retrosplenial cortex.

nature portfolio

Corresponding author(s): Keith Hengen

Last updated by author(s): Jun 17, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	In vivo electrophysiological data was collected chronically with custom micro-electrode and eCube Server (White Matter LLC). Animal behavior was monitored with 15 or 30 fps video recording with WatchTower (White Matter LLC).
Data analysis	Raw data and various permutations were analyzed using a CNN created in TensorFlow. Spike unit activities were sorted using SpikeInterface and MountainSort4. Animal locomotion activities were analyzed from video using DeeplabCut. All other code (Python) required for analyses in the manuscript is available publicly at github.com/hengenlab. Relevant packages and version numbers: open ephys GUI: 0.5.5.3 mountainsort: 4 spikeinterface: 0.100.7 YASA: 0.6.4

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The datasets generated and/or analyzed in this study constitute >10 terabytes of raw neural broadband. The data are stored in a cost efficient manner not immediately accessible to the internet. Data are available upon request.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	n/a
Reporting on race, ethnicity, or other socially relevant groupings	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes for this study are similar to those employed in related studies (PMID: 21525926)
Data exclusions	3 CNN models were excluded that failed to converge. Single units were include if they met standard spike sorting criteria such as high amplitude spiking, stable spike amplitude over time, consistent waveform shapes, and little to no refractory period contamination. Data from 1 implant site was excluded on the grounds of excessive EM noise.
Replication	All experiments were performed with sufficient biological replicates across studied subjects. Additionally, experimental approaches are documented with sufficient detail for precise replication.
Randomization	Subjects were randomly selected before surgical procedures, and were randomized by sex.
Blinding	n\a - there are no groups in this study to be blinded to.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	9 mice (C57BL6/J), 7 female, 2 male. Mean age 220 days.
Wild animals	no wild animals were used in this study.
Reporting on sex	Male (2) and female (7) animals are included in our study. We do not consider sex as a factor in our analyses - but results apply to all animals studied.
Field-collected samples	There were no field collected samples in this study.
Ethics oversight	All procedures involving mice were performed in accordance with protocols approved by the Washington University in Saint Louis Institutional Animal Care and Use Committee, following guidelines described in the US National Institutes of Health Guide for the Care and Use of Laboratory Animals.

Note that full information on the approval of the study protocol must also be provided in the manuscript.